What can machines learn about heart failure? A systematic literature review

International Journal of Data Science and Analytics
https://doi.org/10.1007/s41060-021-00300-1

REVIEW

What can machines learn about heart failure? A systematic literature review

A. Jasinska-Piadlo[1,2] · R. Bond[2] · P. Biglarbeigi[2] · R. Brisk[1,2] · P. Campbell[3] · D. McEneaney[1,4]

Received: 2 July 2021 / Accepted: 3 December 2021

Abstract

This paper presents a systematic literature review with respect to application of data science and machine learning (ML) to heart failure (HF) datasets with the intention of generating both a synthesis of relevant findings and a critical evaluation of approaches, applicability and accuracy in order to inform future work within this field. This paper has a particular intention to consider ways in which the low uptake of ML techniques within clinical practice could be resolved. Literature searches were performed on Scopus (2014-2021), ProQuest and Ovid MEDLINE databases (2014-2021). Search terms included 'heart failure' or 'cardiomyopathy' and 'machine learning', 'data analytics', 'data mining' or 'data science'. 81 out of 1688 articles were included in the review. The majority of studies were retrospective cohort studies. The median size of the patient cohort across all studies was 1944 (min 46, max 93260). The largest patient samples were used in readmission prediction models with the median sample size of 5676 (min. 380, max. 93260). Machine learning methods focused on common HF problems: detection of HF from available dataset, prediction of hospital readmission following index hospitalization, mortality prediction, classification and clustering of HF cohorts into subgroups with distinctive features and response to HF treatment. The most common ML methods used were logistic regression, decision trees, random forest and support vector machines. Information on validation of models was scarce. Based on the authors' affiliations, there was a median 3:1 ratio between IT specialists and clinicians. Over half of studies were co-authored by a collaboration of medical and IT specialists. Approximately 25% of papers were authored solely by IT specialists who did not seek clinical input in data interpretation. The application of ML to datasets, in particular clustering methods, enabled the development of classification models assisting in testing the outcomes of patients with HF. There is, however, a tendency to over-claim the potential usefulness of ML models for clinical practice. The next body of work that is required for this research discipline is the design of randomised controlled trials (RCTs) with the use of ML in an intervention arm in order to prospectively validate these algorithms for real-world clinical utility.

Keywords Heart failure · Machine learning · Data analytics · Data science · Heart failure dataset

1 Introduction

Cardiovascular disease (CVD) is the leading cause of death in the United Kingdom (UK) [1] and worldwide [2]. Advances in pharmacotherapy and invasive strategies have resulted in the increased survival of patients with acute coronary syndromes (ACS), leading to the increased prevalence of heart failure (HF) [3,4]. There are approximately 920,000 people in the UK living with HF and around 200,000 new HF diagnoses each year [5]. HF causes approximately 5% of all emergency adult hospital admissions [6] and it accounts for approximately 2% of total NHS expenditure [6]. HF is a complex condition affecting a wide spectrum of the population [6] and in order to provide credible characteristics of the group of patients with HF, a reliable data collection process

✉ A. Jasinska-Piadlo
jasinska_piadlo-a@ulster.ac.uk

1 Southern Health and Social Care Trust, CVD Research Unit, Craigavon Hospital, 68 Lurgan Road, Portadown BT63 5QQ, Northern Ireland

2 Faculty of Computing, Engineering and the Built Environment, Ulster University, Shore Road, Jordanstown BT37 0QB, Northern Ireland

3 Southern Health and Social Care Trust, Cardiology Department, Craigavon Hospital, 68 Lurgan Road, Portadown BT63 5QQ, Northern Ireland

4 Centre for Advanced Cardiovascular Research, Ulster University, Shore Road, Jordanstown BT37 0QB, Northern Ireland

Published online: 30 December 2021

is needed. The introduction and adoption of electronic health records (EHR) has initiated widespread interest and created opportunities for translational research with respect to cardiovascular health data [7]. The healthcare sector generates 30% of the digital data worldwide [8]. The use of digital clinical data has the potential to transform healthcare systems into "self-learning health systems" [9,10]. In contrast to clinical trials, observational cohort studies based on extracts from digital data and EHR typically do not exclude real-world patients, such as elderly and frail individuals with multiple co-morbidities [11]. Research into health data can change clinical practice and improve patient outcomes, especially for cohorts of patients who would not have been recruited to clinical trials. This was exemplified in Sweden, when a change in antiplatelets prescribing was introduced following discoveries from Swedish Heart Registry - SWEDEHeart [11]. Governments and policymakers increasingly recognise that the healthcare sector is a field where valuable insights can now be uncovered through big data analytics [12]. There are an increasing number of governments that have set out plans for AI in the healthcare sector [12]. In the UK, a policy document published in October 2018 set out the government's vision for the use of technology and digital data within health and care to meet the needs of all NHS users [13]. This included a plan that all the healthcare organisations should have a board-level understanding of how data and technology can drive their services and strategies. The document provided the framework for the NHS to take charge of the digital maturity of its organisations. Collaboration, co-development and iteration between innovators and the NHS were set to become the new norm [13]. In an order to fulfil this requirement, the UK government is funding the Digital Fellow Programme, which has the aim of providing training for clinical staff to develop digital skills [14]. Soon after this policy was made available to the public, the government published 'The Topol Review: Preparing the healthcare workforce to deliver the digital future' which sets out the vision for the NHS in a digital era. The Topol Review, led by cardiologist, geneticist, and digital medicine researcher Dr Eric Topol, explores how to prepare the healthcare workforce to deliver the digital future through education and training. Dr Topol appointed a Review Board and three Expert Advisory Panels [15]. The Topol review argues that data analytics will to be the bread and butter of the future workforce of the NHS. In the light of the increasing incidence of HF and a widespread interest in ML application to health data, we present a literature review of studies using ML to analyse HF datasets. The aim of this review is to learn how ML can complement current clinical practice and management of patients with HF.

2 Outline of the paper

This systematic literature review is comprised of eight sections. Section 1, Introduction, outlines the rationale for the review. Section 3 provides a summary of the previous literature reviews undertaken in the field of HF and ML. Section 4, Material and Methods, provides the search criteria, scope notes for the search criteria and inclusion and exclusion criteria. In Sect. 5, Results, we describe the common HF problems addressed by ML in the reviewed papers, general characteristics of the studies, such as sources of the datasets, sample size, types of variables used in studies, management of missing data, ML algorithms used and their performance. This section considers how gaps in the field have been addressed in recent years and the impact of ML on the progress of HF problem solving. In Sect. 6, Discussion, We discuss the role of predictive models in international HF guidelines. In Sect. 7, Gaps and Research Opportunities, we discuss the directions for further research in ML and HF, as well as the need for development of models using modern HF patient cohorts and the standards for the reporting of studies using ML and AI in healthcare data. Section 8, Conclusions, provides a summary of most important elements of the review and outlines the limitations of the review.

3 Previous systematic literature reviews on ML and HF dataset

There are previous systematic reviews presenting studies on ML and HF datasets. Rahimi et al. (2014) reviewed ML methods, discrimination, calibration and model validation methods of studies from 1995 to 2013 [16]. They concluded that risk prediction models have low uptake amongst clinicians [16]. In support of this thesis, they cited the Postal Survey of Physicians attitudes to implement cardiovascular prediction rules [17]. Tripolity et al. (2017) reviewed HF classification models from 2000-2016 [18]. They observed that in most cases researchers focused on two or three-class classification problems of HF severity [18] even though etiology and symptoms of HF are more complex and can not be fully addressed by answering dichotomous questions. Alba et al. (2013) reviewed the mortality prediction models of ambulatory patients with HF [19]. They observed that out of 32 studies, only 5 studies (15%) were validated in an independent cohort of patients [19]. The lack of validation on external datasets made it impossible to evaluate how well the models would generalise in a real-world clinical setting. Mahajan et al. (2018) reviewed predictive models for identifying the risk of readmission after index hospitalization for HF and suggested that more work needs to be done for calibration, external validation, and deployment of predictive models to ensure suitability for clinical use [20]. Bazoukis

et al. (2020) included studies using data from heterogeneous sources - clinical trials, data from cardiopulmonary exercise stress tests, left ventricular assist device (LVAD) and cardiac resynchronisation therapy (CRTP) [21]. They performed a quality assessment of the ML studies using a novel score proposed by Qiao [22]. They concluded that 'at the moment ML could not replace clinical cardiologists', which was preceded by disclaimer that analysis of healthcare data with ML techniques still act as an auxiliary decisional role [21].

4 Methods

4.1 Identification of studies and literature searches

Literature searches were performed in November 2019 and updated in February 2021. This literature review follows a systematic review methodology whereby the PRISMA framework was used to evaluate the suitability of studies for inclusion within the review. Figure 1 shows the flow diagram of the identification process for articles included in this review. The initial search identified 2679 articles published between 1/2014 and 2/2021, of which 1688 studies remained after removing duplicates. Out of 1688 studies, 1497 were excluded based on the title and abstract screen. 191 studies were considered for full text analysis. 110 studies were excluded during full-text screening. Eighty-one studies met the inclusion criteria and were included in the qualitative and quantitative analysis

SCOPUS, ProQuest and MEDLINE Ovid databases were searched using the following terms - search terms included ("heart failure" OR cardiomyopathy/ies OR "cardiac oedema" OR "paroxysmal dyspnoea") AND ("machine learning" OR "data mining" OR "data analytics" OR "data science"). The scope note for the term "heart failure" includes: cardiac failure, congestive heart failure, decompensation, heart decompensation, congestive heart failure, left sided heart failure, left-sided heart failure, right sided heart failure, right-sided and left sided heart failure, myocardial failure, heart failure with reduced ejection fraction (HFrEF), heart failure with preserved ejection fraction (HFpEF). The scope note for "cardiomyopathy/ies" includes: a group of diseases in which the dominant feature is the involvement of the cardiac muscle itself; cardiomyopathies are classified according to their predominant pathophysiological features (dilated cardiomyopathy, hypertrophic cardiomyopathy, restrictive cardiomyopathy) or their etiological/pathological factors alcoholic cardiomyopathy, endocardial fibroelastosis, primary and secondary cardiomyopathies, primary and secondary myocardial disease.

The SCOPUS database assigns the following definition to "machine learning": a type of artificial intelligence (AI) that enables computers to independently initiate and execute learning when exposed to new data. The scope note for "machine learning" includes: machine learning and transfer learning. Data science is defined as "an interdisciplinary field involving processes, theories, concepts, tools, and technologies, that enable the review, analysis, and extraction of valuable knowledge and information from structured and unstructured (raw) data" [24,25]. The scope note includes data analytics, data driven science and data science. A search limit was applied to include original articles and conference papers published in English. The titles and abstracts of full-text articles were screened for suitability after applying inclusion and exclusion criteria.

4.2 Inclusion and exclusion criteria

Inclusion and exclusion criteria were agreed between clinicians and data scientists. Clinicians defined the most relevant aspects of heart failure (HF) identification, detection and diagnosis in the context of the application of machine learning (ML) to electronic health records. Selected studies used ML techniques to analyse HF datasets to predict the following outcomes: worsening of a clinical condition, readmission to hospital, onset of illness, classification of HF stage according to symptoms reported by patients, classification based on the HF etiology, HF mortality, response to introduced HF treatment. Included studies were published from 2014 until February 2021 (inclusive). Included studies applied ML to the range of medical datasets of HF patients, i.e. electronic health records (EHR), datasets from primary or secondary care, open access HF data sets and repositories with data from patients with cardiovascular disease. We excluded studies whose primary focus was the analysis of: ECG signals of HF patients, echocardiographic video loops, biobank datasets, image repositories and image signals from CMR and PET, histology and pathology datasets from cardiomyopathy cases, the ECG signals and results of the cardiopulmonary exercise stress test (CPEST), data from cardiac resynchronisation therapy devices (CRT, CRTP, CRTD), data exclusively from extra-corporal life support (ECLS) like blood parameters. Studies were excluded if they exclusively included mobile health and telehealth datasets of HF patients, health claim databases and healthcare cost analysis related to HF cohorts. Studies using natural language processing (NLP) and text mining techniques as the only means to interrogate the HF dataset were excluded. We excluded studies which focused on the analysis of health record access patterns whilst not analysing the health records themselves. Theses, reviews, book chapters, editorials, letters, conference abstracts without full text and non-English articles were excluded. A data extraction table was used to record features of interest from each study, including study quality indicators. Predefined criteria and its details are listed in a supplementary material.

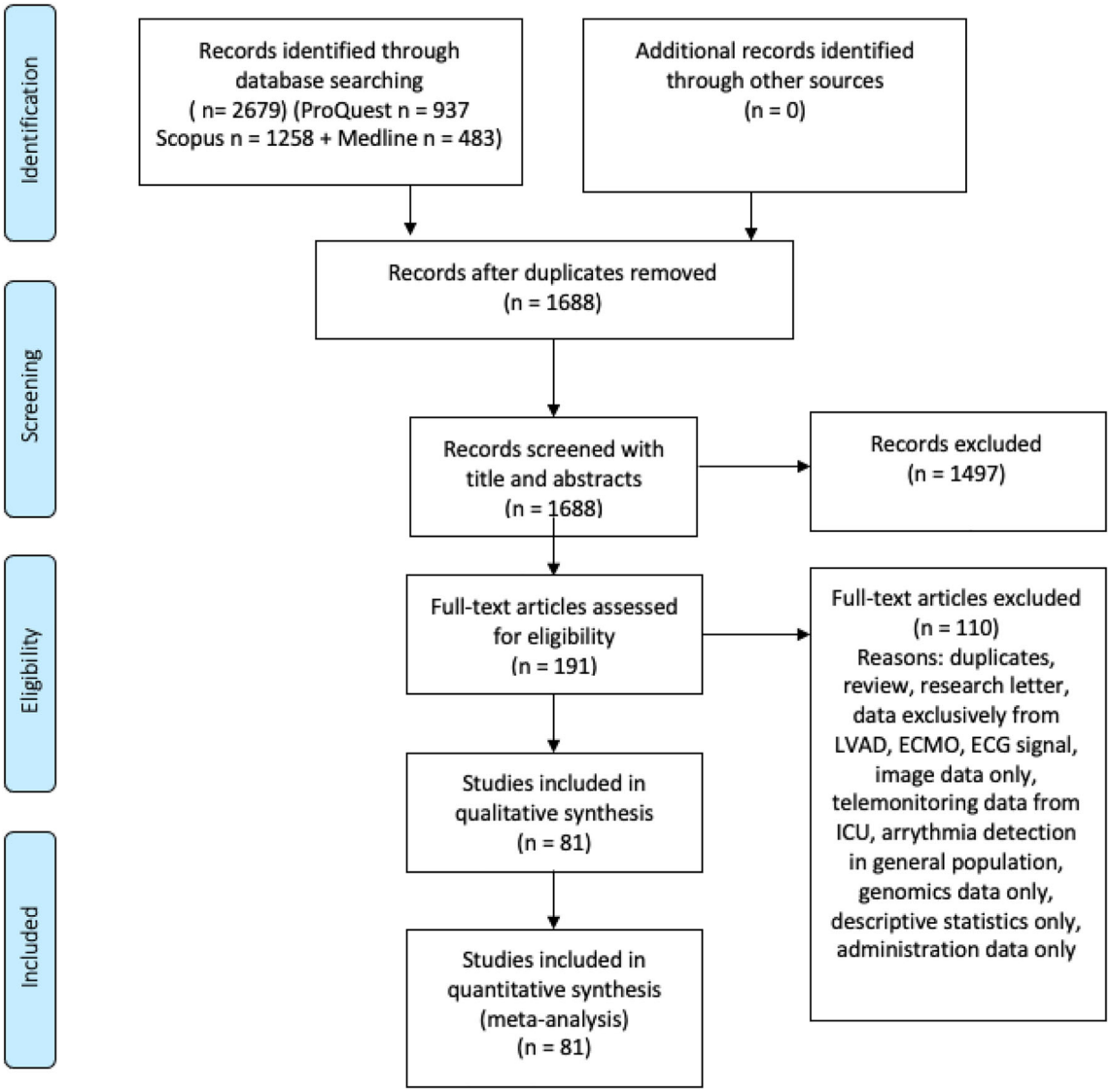

Fig. 1 PRISMA flow diagram of the identification process for articles included in this review. PRISMA = preferred reporting items for systematic reviews and meta-analyses; Adapted From: Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. [23]

5 Results

5.1 Common HF problems addressed by ML

At the outset of the Results section, it is useful to outline common issues in HF that require attention of healthcare providers. For years clinicians have tended to focus upon providing effective care to HF patients, however the interest of data scientists to provide solutions to this domain creates the opportunity to improve HF services by employing ML to address common HF problems. HF is a complex clinical syndrome and due to the shear volume of specific problems associated with this syndrome the inclusion criteria were employed to narrow the focus of studies to those working with clinical and administrative HF datasets. This review is focused upon the ways in which EHR, clinical and administrative data can reveal correlations within patient data that inform both research and clinical practice. Therefore this review will not focus upon studies analysing ML application to electrocardiogram (ECG) nor cardiac imaging. The extensive research into electrocardiogram (ECG) signal processing and ECG interpretation has been recently synthesised in the state-of-the-art systematic review of application of deep learning to the ECG [26]. Somani et al. (2021) provide an overview of deep learning application to ECGs, its benefits, limitations as well as future areas for improvement [26]. We recognise however that areas of cardiology where ML and AI solutions have been successfully applied to ECG analysis and cardiac imaging which has resulted with the FDA approval in clinical settings. This is exemplified by soft-

Table 1 Outcomes examined in the studies included in the literature review. N - number of studies, % out of 81 studies. *Other - modelling of the response to HF treatment, reduction of HF data dimension

Outcomes of studies	N/%	Citations to studies
Detection of HF onset	25 (31%)	[29–53]
Mortality prediction	20 (26%)	[54–73]
Prediction of readmission to hospital	18 (21%)	[74–91]
Classification of HF (according to NYHA class or aetiology) or clustering	11 (13%)	[92–102]
Other: *	5 (6%)	[35,62,103–106]

ware to automated ECG interpretation and reporting, deep learning enabled heart function analysis of cardiac magnetic resonance (CMR) images and ECHO imaging [27]. Cardiac imaging is a broad and well established sub-speciality of cardiology. As with the case of ECG, we refer readers to the systematic review by Dey et al. (2019) which provides insight into AI application to cardiac imaging [28]. Below we list common problems within the heart failure domain addressed by ML application to electronic health records:

1. HF detection and diagnosis from electronic health records or administrative data.
2. Prediction of HF readmission to hospital (30-day, 60-day, 3-month, 6-month readmission since the index hospitalisation). Studies within this domain problem focused upon grouping patients according to the readmission risk. The search for the accurate prediction of HF readmission is likely influenced by incentive programs in the USA that record readmission rates in the annual payment update by the insurance company Medicare.
3. HF mortality prediction based on patients' current clinical status, results of routine blood tests (biomarkers), non-invasive cardiac imaging - echocardiogram (ECHO), magnetic resonance imaging (MRI), myocardial perfusion scan (MPS) and invasive tests such as cardiac catheterisation to assess pressures in heart chambers, and coronary angiogram and intravascular ultrasound (IVUS) - to assess coronary artery plaque volume and composition.
4. HF classification applied in line with already known HF categories, based on patients' symptoms such as New York Heart Association (NYHA) and based on the type of HF derived from international HF guidelines such as classification according to the functional assessment of left ventricle by calculating ejection fraction (LVEF). Distinct types of HF would include HF with reduced EF (HFrEF), HF with mid range EF (HFmrEF) and HF with preserved EF (HFpEF).
5. HF classification using unsupervised ML methods in order to discover new phenotypes of HF patients, not described today based on the above criteria.
6. Prediction of the outcomes and response to invasive therapies like implantable cardiac devices: implantable cardioverter defibrilator (ICD), cardiac resynchronisation therapy (CRTP/D) and left ventricle assisted devices (LVAD).

In this review, we grouped studies according to the specific HF problem which they aimed to address. Based on this approach we produced a taxonomy of common HF problems investigated by ML applied to HF datasets. Table 1 presents references to studies grouped according to the domain problem that they attempt to solve. In Fig. 2, we present a taxonomy of the common HF problems and which ML methods were applied to solve those problems.

5.2 General study characteristics: variables, source and size of datasets used.

All 81 studies were retrospective, observational or cross-sectional cohort studies. HF datasets comprised a mixture of continuous and discrete variables. Most studies used either structured or unstructured data, though some studies integrated both and described it as a novel approach in their data analytics. Clinical and administrative features were considered, as listed in Table 2. The most common source of HF patient data used for analysis in reviewed studies was an extract from Electronic Health Records (EHR) of large clinical centers, university hospitals, district general hospitals or community health centers. Several studies used registry data as in a case of Swedish national cardiovascular registry [55], acute HF registry enrolling patients from 10 hospitals across Korea [57], data from BIOSTAT-CHF project, which enrolled patients from 69 centres in 11 European countries between 2010 and 2012 to determine profiles of patients with HF that do not respond to recommended therapies [59], or as in a case of Frizell at al. (2017) study dataset was obtained by linking patients from the "Get With the Guidelines" Heart Failure registry with Medicare data from 289 hospitals in USA [79]. Blackstone et al. (2018) [73] aimed to develop a decision aid to aggregate adverse events in heart transplant and measure end organ function to inform clinical decision making. They used the Electronic Data Interface

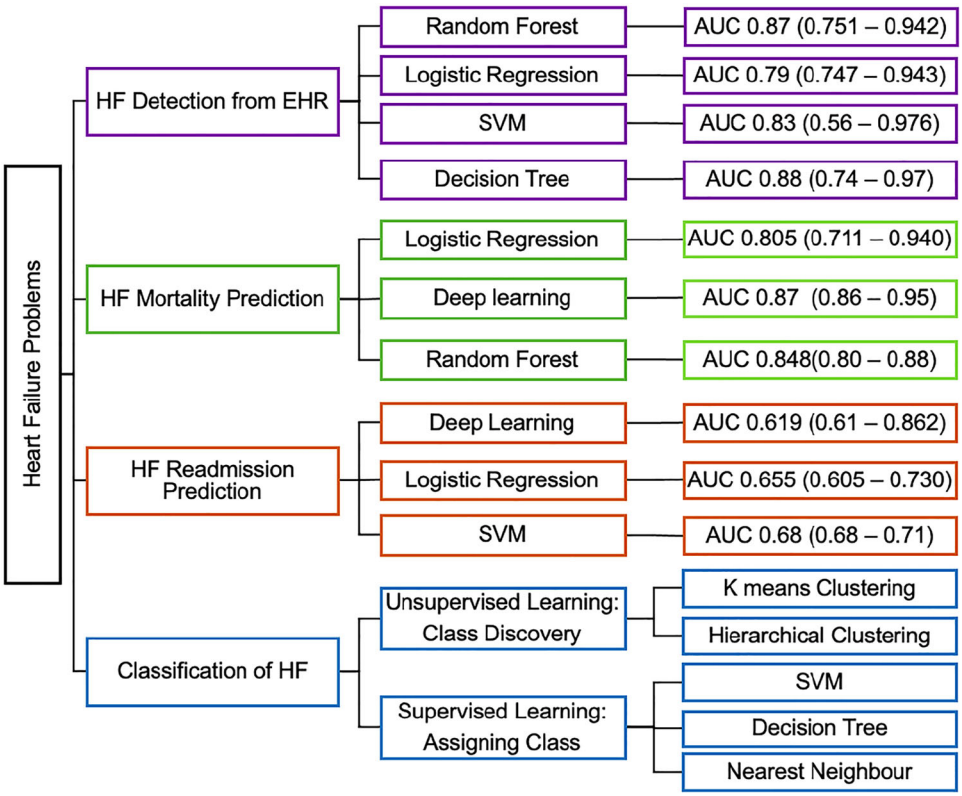

Fig. 2 Taxonomy of the main problems within the HF domain. This Figure presents the common HF problems and most commonly used ML algorithms in addressing these specific HF domain problems. HF problems are colour coded and ML methods used in solving these problems are colour coded and listed in order of the most frequently used method for each particular domain problem. For each ML method used in the specific domain problem, there is a median AUC provided, with the performance range (min-max) achieved by ML model applied to the specific problem

for Transplant (EDIT) database, updated by transplant coordinators during the course of clinical care and data from The Cardiovascular Information Registry (CVIR), a prospective registry of all cardiovascular procedures performed in Cleveland Clinic, Ohio, USA. These data were supplemented with queries from EHR to resolve inconsistencies and impute the incomplete data.

Several studies however used open-source cardiovascular data available from data repositories available in a public domain. Below we provide the web address to the open-source cardiovascular data repositories used in papers cited in this review:

1. Cleveland Clinic Foundation Heart Disease Data Set from University of California Irvine (UCI) Machine Learning Repository available on https://archive.ics.uci.edu/ml/datasets/heart+disease
2. MIMIC - III Medical Information Mart for Intensive Care from Beth Israel Deaconess Medical Center in Boston, Massachusetts available on https://github.com/MIT-LCP/mimic-website

Eleven studies [29–31] [34,37,38,50,84,92,105,107] used Heart Disease Data Set from the ML Repository of the Cleveland Clinic Foundation and University of California, Irvine (UCI) [108]. This dataset was used by 8 out of 10 studies to develop a model to predict the onset of HF. Two studies developed the HF classification model. Apart from Cleveland Clinic repository, real-world datasets from Microsoft Azure research platform [59] were used and the publicly available MIMIC-III benchmark datasets from critical care databases.

In case of multimodal data, including live and still imaging data and ECG signals, they are often stored on multiple platforms and due to the nature of the data (coded millions of pixels of still images, video loops). The process of integrating this information with clinical information readily available in HER poses a challenge. Patient data are fragmented and scattered across multiple silos, which would require integration prior to application of advanced data analytics.

In Table 3, we list most commonly used variables utilised by cited authors in their final models. The most commonly used variables in mortality prediction models were (in order of most frequently used) (1) left ventricular ejection fraction (LVEF), (2) the presence of other clinical conditions (comorbidities), (3) age and (4) renal function (as measured by serum creatinine level). In the group of HF classification into different types of HF or different stages of HF, the most commonly used variables were (1) the presence of hypertension, (2) age, (3) gender, (4) presence of coronary artery disease, (5) blood tests (wide range of blood tests), (6) renal function tests (serum creatinine level and sodium level). In the group of HF onset prediction, the most commonly used variables were (1) age, (2) presence of diabetes and (3) hypertension. The number of variables used across all studies ranged from

Table 2 Examples of discrete and continuous data used in studies presented in the systematic literature review

Discrete data	Continuous data
Demographics	*Physical examination:*
Age / Sex / Gender: Female/ Male	Pulse Rate (beats per minute)
Race: White, African American, Hispanic, American Indian,Native Asian	Respiratory Rate (breath per minute)
Medicare Insurance / Medi-Cal Insurance	Systolic Pressure in mmHg
Contacts with Healthcare/Care management:	Body Mass Index (BMI)
Discharge to Skilled Nursing Facility	Body Surface Area (BSA)
Missed Clinic Visits in Prior Year	*Blood tests*
ED and O/P Visits in Prior Year	Biochemistry:
Admission in Previous Year	Serum B- Natriuretic Peptide (pg/mL)
In-admission Telemetry Monitoring	Glucose (mg/dL),
Clinical data	Fasting blood glucose >120 mg/dl
Symptoms:	Serum Creatinine (mg/dL)
Types of chest pain, Breathlessness as per NYHA class	Urea (mg/dL)
Past Medical History:	Serum Albumin (g/dL)
Ischemic Heart Disease	Cholesterol level
Previous Myocardial Infarction	Serum Sodium and Potassium (mEq/L)
Previous Heart Failure	Haematology:
Type of Cardiomyopathy	Haemoglobin (g/dL), Haematocrit (%)
Coronary Artery Disease	*Additional tests*
Valvular Heart Disease	ECG (recorded at rest) features
Arrhythmias	heart rhythm: sinus rhythm atrial fibrillation
Cerebrovascular Disease/Stroke/TIA	ORS width - broad or narrow
Vascular/Circulatory Disease	Exercise Stress Test (EST):
Diabetes type I and II	MPHR - maximum predicted heart rate
Renal Disease or ESRD on Dialysis	EST induced angina, ST segment depression, downslope of ST segment or upslope of ST segment
Chronic Lung Disease/COPD/Asthma	ECHO features
Metastatic Cancer of solid organ or Acute Leukemia	LVEF in % (left ventricular Ejection Fraction)
Severe haematological disorder	Right ventricular systolic pressure
Liver Disease	Pulmonary artery mean pressure
Mental Disorder(s)	Chest XRay features:
Medication History	Lung fields
Social History: Alcohol Abuse, Drug Abuse, Protein Caloric Malnutrition, Functional Disabilities	Cardiomegaly

8 to 4205. Figure 4 shows the number of variables used in models applied to specific outcomes groups. The majority of authors reported the exact number of variables included in their model. A small minority provided neither the total number of variables nor any clear description of the variables used.

Figure 3 shows the median size of the patient cohort examined in studies. The largest patient samples were used in readmission prediction models with the median sample size of 5676 (min. 380, max. 93260), followed by mortality prediction models with a median sample size of 5044 (min. 95, max. 44886), followed by studies focused on classification of HF with median sample size of 853 (min 162, max 41713) and the smallest samples were used in HF prediction models with median sample of 439 (min 46, max 67697).

5.3 Dimensionality of datasets

The healthcare data are highly dimensional. Several studies reviewed aimed to reduce data dimensionality [50,105] in order to improve the performance of their algorithm. Models trained on datasets with many features and limited number

Table 3 Most commonly used variables and median numbers of variables used in final models in cited studies; all grouped by outcomes of the studies

Study outcome	Median number of variables	Most common variables
Detection of HF	14 (range: 13 - 1823)	Age, presence of: diabetes, hypertension
HF mortality prediction	45 (range: 8 - 1302)	LVEF, comorbidities, age, renal function tests (creatinine, urea)
HF classification	55 (range: 11 - 400)	Hypertension, age, gender, coronary artery disease, blood tests, renal function tests
Prediction of HF readmission	56 (range: 16 - 4205)	Age, blood tests, comorbidities

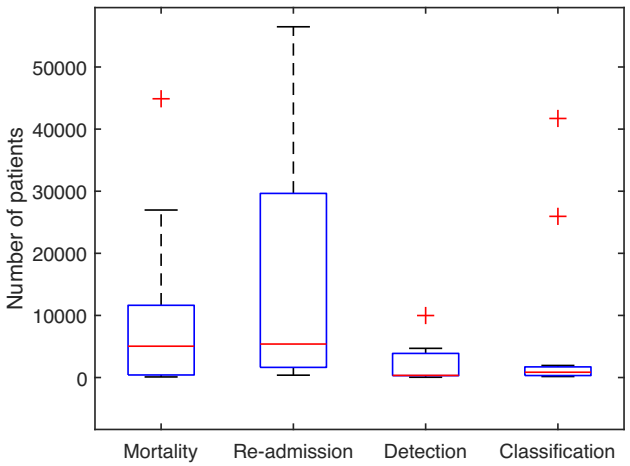

Fig. 3 Patients' sample size used in ML studies grouped by examined outcomes. (On each box, the central mark indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points considered outliers, and the outliers are plotted individually using the '+' symbol. For better visualisation effect, 3 studies have been removed from the plot. Studies working on 93260 (HF Readmission), 67697 (HF Prediction) and 41713 (HF Classification) patients are not included in the figure.)

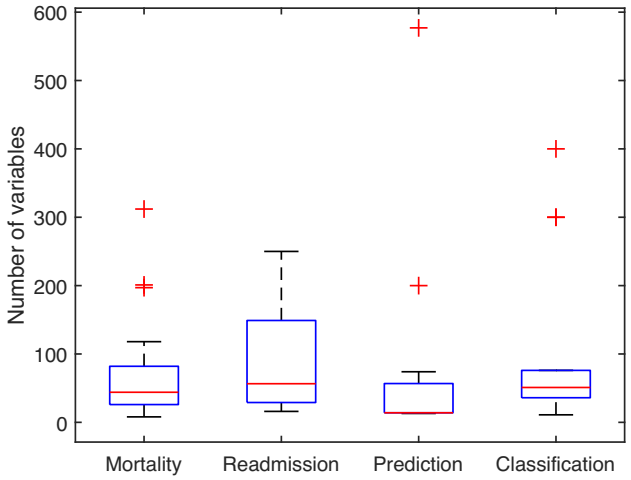

Fig. 4 Number of variables used in ML studies, grouped by outcomes. (On each box, the central mark indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points considered outliers, and the outliers are plotted individually using the '+' symbol. For better visualisation effect, 4 studies were removed from this plot. Studies using 4205 (HF Readmission prediction), 1823 (Detection of HF), 1302 (HF Mortality prediction) and 939 (HF Readmission prediction) variables in their model.)

of recorded observations for each variable are exposed to a risk of over-fitting. This leads to a reduction in the model performance when it is applied to an external dataset. In Fig. 6 we illustrated the relationship between sample sizes of all studies and number of variables used. Studies with sample size less than 10,000 patients achieved median AUC 0.86, whereas studies with sample sizes greater than 10,000 patients achieved median AUC of 0.814. We correlated the ratio between sample size and the number variables with the study performance as shown in Fig. 5c. While we would expect studies with smaller sample sizes to use a fewer number of variables in order to improve algorithm performance, we did not observe much consideration of the dimensionality aspect in reviewed studies. Surprisingly there was no correlation observed between the algorithm performance and sample size to number of variables ratio. Figure 7 shows the ratio of the number of variables to the number of patients included in the study grouped by the HF problem that they address.

5.4 Handling of missing data

Most real-world datasets contain missing values. This can cause issues for a number of ML methods [109]. The percentage of missing values differed between studies. In many cases, the variables with missing values were removed from the dataset. For example, Sideris et al. (2016) excluded features corresponding to patient weight from their analysis because 97% of these values were missing [84]. They also removed information about the medical specialty of the admitting physician because this information was missing in

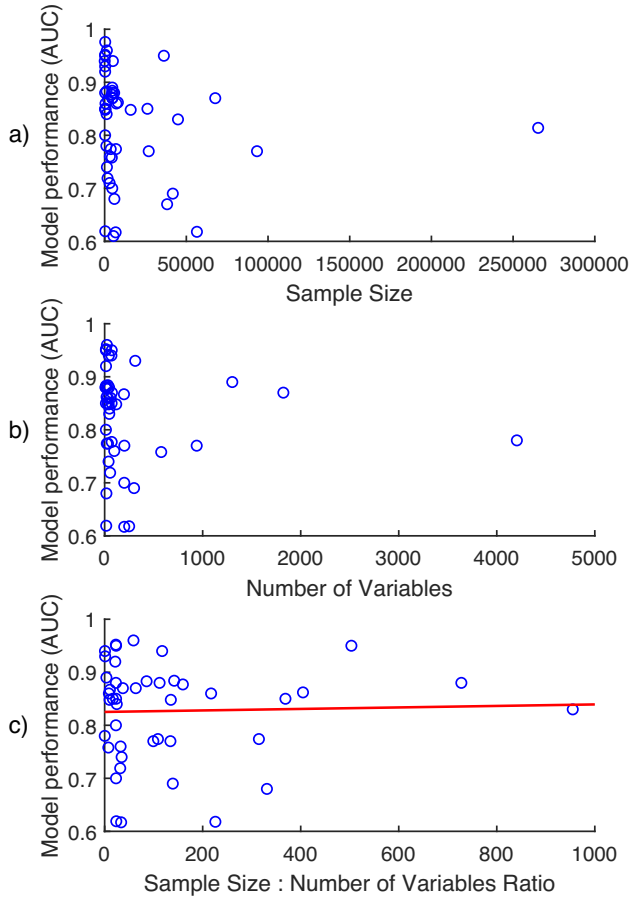

Fig. 5 Representation of the ML model performance in relation to number of variables used, sample size and the ratio between sample size and number of variables. **a** Number of variables used in a study vs ML model performance. **b** Sample size used in a study vs ML model performance. **c** Ratio of the sample size to the number of variables used in the ML algorithm vs ML model performance. Least square regression line indicates negative correlation between model performance and the number of variables per patient in the sample cohort (R value 0.03)

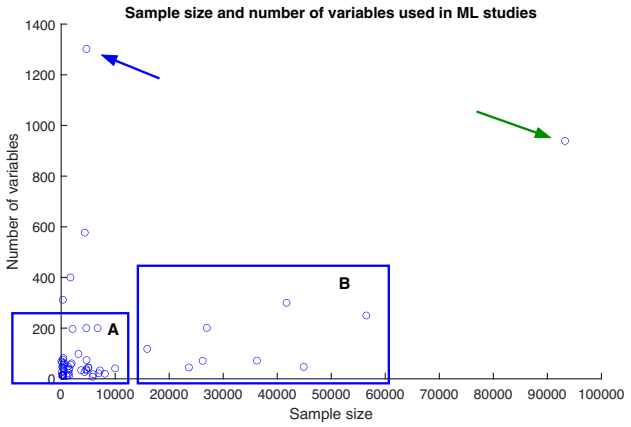

Fig. 6 Scatter plot illustrating sample size used in the ML study and variables tested in the same study. For better visualisation effect 2 studies using 4205 and 1823 variables in their model have been removed from this plot. Studies with smaller sample sizes, less than 10,000 patients and variable size less than 200 (Cluster A), achieved median AUC 0.86, whereas studies with sample sizes greater than 10,000 patients and variables number less than 400 (Cluster B) had median AUC of 0.814. Blue arrow indicates a study which used 1302 variables. This study used multi-view ensemble learning based on empirical kernel mapping to predict HF mortality and achieved AUC of 0.89. Green arrow indicates a study, that used 939 variables to train neural network model to predict HF hospital readmission. This model achieved AUC of 0.77

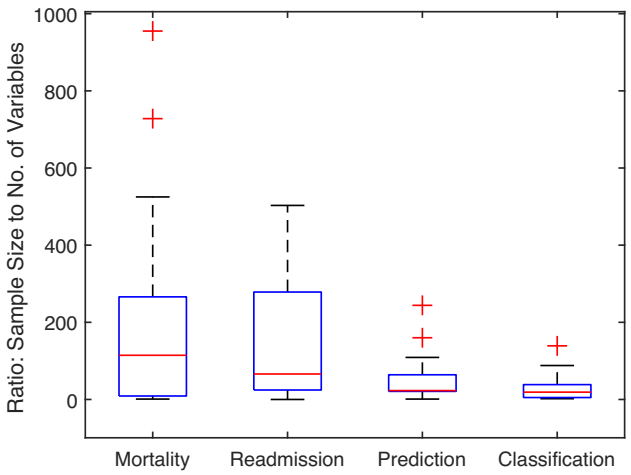

Fig. 7 Box plots illustrating ratio between sample size and number of variables used in reviewed studies. (On each box, the central mark indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points considered outliers, and the outliers are plotted individually using the '+' symbol.)

49% of cases. Similarly, information about the payer code was removed from the final dataset, because it was unknown for 39.5% of patients [84]. For Ahmad at al. (2018), over 20% of missing data for a specific variable was enough to exclude this variable from the analysis [55]. According to the authors, the most likely missing variables were laboratory values. They observed an impact on prognostication and clustering when variables of known prognostic value were missing, such as B-natriuretic peptide (BNP) serum levels [55]. Chu et al. (2020) excluded patient samples with more than 30% of missing values from the analysis [68]. The Cleveland heart disease dataset contains records of 303 patients with 76 attributes. From this dataset usually 6 patients are removed in order to perform analysis [29–31,34,37,38,50,84,84,92,105,107]. Researchers worked on 297 records without any missing values and tended to use 13 key attributes [31]. A variety of approaches were used to input missing data. Schrub et al. (2020) used the SVDimpute function within the impute package in R to impute the missing data [100]. They report that the percentage of missing values ranged from 0% to 28% [100]. Kwon et al. (2019) in an acute HF mortality model also replaced missing data using data imputation [57]. They used patients' most recently recorded values of the missing feature to complete variables

such as the vital signs, demographics, biochemistry results or heart scan features that were available in patients' notes. Using this method Kwon et al. (2019), created a training dataset consisting of 12,654 datasets amplified from 2165 patients. As a result, many training datasets were generated based on data sourced from the records of only one patient. In the opinion of Kwon et al. (2019) this dataset was sufficient for developing a deep-learning model to predict HF mortality [57]. Their deep neural network algorithm achieved an AUC score of 0.880 [57]. Nonetheless, this model outperformed both the American Cardiac Association "Get With The Guidelines' (GTWG-HF) risk model and The Meta-Analysis Global Group in Chronic Heart Failure (MAGGIC) risk model for predicting 12- and 36-month mortality [57].

5.5 Overview of algorithms

Table 4 presents all ML algorithms that have been employed within the reviewed studies. Some studies used more than one ML method; hence, the total number of methods used is larger than the number of studies included in the literature review. Most commonly We observed an increase in the use of deep learning in tackling HF problems. Neural networks performed very well in predicting HF mortality or detecting HF from EHR. In the majority of the studies using neural networks, they outperformed traditional ML methods [56,68,70,72]. Hence, deep learning (DL) could potentially improve the accuracy of HF classification, detection, and outcome prediction. However, DL is typically applied when using image data (e.g. echocardiograms, CMR, CT) or ECG data (e.g. using CNNs or RNNs) but can also be used for other datasets, including text analysis. Moreover, the problem with DL models is that they lack explainability, hence there is a trade off between accuracy and explainability. Therefore, the adoption of DL in clinical practice can be challenging given the need for accountability. Hence, one could argue that traditional ML could be preferred given that some of these techniques can provide an insight into the rationale and logic behind the computer's recommendation. Nevertheless, there is ongoing research that is investigating methods to explain DL models, for example using attention maps [110] which allow the user to see what features the DL algorithm focused on just before it produced its classification.

5.6 Algorithms performance

The comparison of algorithmic performance in studies considered within the review is challenging due to each study's performance being reported using variable ML evaluation metrics. Firstly, studies that used data from heterogeneous populations and models, have been trained on populations with different sample distributions and characteristics. Secondly, studies were set to predict a number of different outcomes. Thirdly, authors used various numbers of variables and some authors introduced a features selection process to identify most accurate predictors to be used in the final predictive model. The described heterogeneity of study characteristics prevents us from performing quantitative comparison of achieved performance by reviewed models. Figure 2 presents a taxonomy of the common HF problems and provides information on the range of the performance of ML methods applied to the specific problem. Figure 2 presents a taxonomy of the common HF problems and provides information on the range of the performance of ML methods applied to the specific problem. In studies presented in this review, the authors used various performance metrics. Area Under the Receiver Operating Characteristics (AUROC or C-statistic) were the most common; followed by sensitivity, specificity, accuracy and F1 score. The authors typically compared their model's performance to previously published studies. There was a tendency to point out that previous models achieved a lesser performance, even when their model only performed better by a fraction under the ROC curve. In studies predicting HF onset and detection of HF, the most commonly used algorithm was random forest (RF), followed by logistic regression, SVM and decision tree methods. Best median AUC was achieved by Decision Tree with median AUC 0.88, followed by RF - median AUC 0.87, and SVM - median AUC 0.83. Logistic Regression was the most commonly used method in mortality prediction studies, followed by RF and deep learning (neural networks). The highest median AUC was achieved by neural networks at 0.870, followed by random forest with median AUC of 0.848 and logistic regression with a median AUC of 0.805. In studies predicting the readmission to hospital, the most commonly used method was deep learning, followed by logistic regression, SVM and RF. The best median AUC was achieved by SVM -median AUC 0.68, followed by logistic regression - AUC 0.655 and deep learning - AUC 0.619. Various methods of clustering were used in studies performing the classification of HF, i.e. K-means Clustering, K-Nearest Neighbour method, hierarchical clustering, SVM, decision tree. Despite the wide use of modern ML methods, logistic regression was still the most common method used in predictive models. Nevertheless, ML models showed improved performance over models using traditional statistical methods. In general, the use of a ML models employing variety of additional variables improve HF mortality risk prediction when compared with conventional approaches like regression models alone [65].

5.7 Addressing previous gaps in ML

Research into ML application to HF datasets evolved and brought some needed solutions to previously identified gaps

Table 4 Types of Machine Learning methods used in cited studies. N (%) number of models (% out of 81 studies)

ML alghorithms	N (%) number of (models (% out of 81 studies)
Supervised ML methods	
Logistic regression (LR) (15) (including Boosted LR, Regularised LR, Knowledge Driven Scalable Orthogonal Regression, Spike-and-slab regression, Multivariable regression, Stepwise LR, Ensemble LR)	22 (26%)
Decision Tree (including Decision Tree ID3 (10), Boosted Decision Tree (2), Boosted Regression Tree (3)	15 (18%)
Random Forest	14 (17%)
Support Vector Machine (SVM)	12 (14%)
Naive Bayes (NB) (including NB (5), Tree Augmented NB (2), Gaussian NB (1))	8 (9%)
Deep learning	
Neural Networks (NN) (including Recurrent NN (7), Convolutional NN (2), Deep NN (1))	10 (12%)
Unsupervised ML	
Clustering Methods (including k nearest neighbour (6), k-means clustering (3), hierarchical clustering (1))	9 (11%)
Selection Operator Models	1
Feature Rankin Analysis	1

in ML methods. Below we describe how recent studies addressed those unmet needs:

1. Using synthetic data. Even though there is still wider need for improvement of validation methods, Xiao et al. (2018) [82] successfully used synthetic data as a validation cohort in their readmission prediction study. They used synthetic data as a benchmark for reproducing experimental results. After generating 3000 synthetic patients, they used synthetic data of 500 patients for validation, and another 500 for testing.
2. Using algorithms to tackle missing data. It has been noted that the omission of variables with a considerable number of missing values should not be a routine approach in ML studies [111]. There has been an increase in using algorithms to address missing values such as multiple imputation techniques, rather than ignoring variables from analysis. Blackstone et al. (2018) [73] used 5-fold multiple imputation using a Markov Chain Monte Carlo technique to obtain final parameter estimates and a variance-covariance matrix to deal with missing variables. Mahajan et al. (2017) [78] used multiple imputation by chained equations re-sampled over five imputed datasets to fill in missing values for used variables. Jiang et al. (2019) [83] used simply mean imputation to impute missing fields of important variables. The above-described methods of dealing with missing data allowed researchers to deal with the most difficult aspect of real-world data - missing values.
3. Feature selection techniques. Another difficult aspect of real world data - high data dimensionality - was dealt with successfully, in number of studies. Wang et al. [67] in their HF mortality prediction systems applied Orthogonal Relief (OR) algorithm to remove irrelevant and redundant features from the dataset. This approach significantly reduced the dimensionality of data and allowed researchers to successfully use the Dynamic Radius Means classification algorithm to predict mortality from HF.
4. Improved ML execution time. Haq et al. [29] used feature selection algorithms in their classification algorithm and not only the accuracy of the model increased but the execution time of the diagnosis system was significantly reduced.
5. Use of ensemble models. In our review we observed a new trend to use multi-model predictive methods. This multi-model architecture can provide better accuracy than best model approach and it has been used successfully by research groups. Priyanka et al. (2016) [33] and Cheung (2018) [76] proposed a novel hybrid model bridging multi-task deep learning and K-nearest neighbors (KNN) for individualised treatment outcome estimation in HF patients. This model achieved F1 score of 0.796 and outperformed state-of-the-art ML predictive models.

5.8 ML impact on advancing HF problem-solving

In this review we highlight exemplary work presenting how ML and data analysis impacted HF problem-solving and have advanced the knowledge in this domain.

1. Personalised medicine. Probably the most important advancement is an important role of ML in executing problems of the personalised medicine. Ahmad et al. (2018) [55], who worked on the SWEDEHeart registry data, presented potential clinical implications of their HF clustering study. They concluded that if clustering into distinctive HF phenotypes was embedded within EHR, they would ultimately create a self-learning healthcare system that could suggest a personalised therapy for specific needs of individual patients [55]. Another good example is the use of deep learning in identifying the important factors which increase patient mortality. This could potentially help clinicians to augment their clinical decision and review planned interventions for HF patients [61].
2. Improving clinical trial design. Multiple studies using ML identified features carrying high predictive value in HF course, confirming what RCTs in HF with reduced ejection fraction (HFrEF) showed to date. There is an opportunity that advance data analytics of HF datasets will lead to identification of new features in the HF pathology, that could be targeted during future clinical trials. ML based clustering gives unique opportunity to identify distinct phenogroups of various types of HF, caused by different etiologies. HF with preserved EF (HFpEF) is a common type of HF, still poorly understood. Gu et al. [99] showed that there are 3 distinct cohorts of patients within a group of patients with HF with preserved ejection fraction (HFpEF). Those cohorts are characterised by significant differences in comorbidity burden, underlying cardiac abnormalities, and long-term prognosis. They observed that beta-blockers or ACEi/ARB therapy was associated with a lower risk of adverse events in specific HFpEF phenogroups [99]. Again, this finding should be explored further and considered for validation in prospective clinical trials. Ahmad et al. [55] suggested that ML has a significant role to play in improving clinical trial design and execution.
3. Healthcare data quality and data integration. Liu et al. (2019) [91] proposed methodology and process on constructing large-scale patient cohorts which allowed to form the basis for effective clinical case review and efficient epidemiological analysis of complex medical conditions. Ben-Asouli et al. (2019) [56] recommended that policymakers should allocate resources to promote projects that bring big data analytics closer to clinical practice. They concluded that there is an opportunity to improve patients' outcomes by investing in comprehensive, integrated health IT systems and projects aimed at simplifying ML to clinical teams [56].

5.9 Authors affiliation and input from clinicians into ML studies

This review included papers from 514 authors. Of these, 297 authors were affiliated with either computer science, informatics, statistics or a related area. A total of 213 authors were affiliated with a medical centre and had MD or MBBS titles. There was an average 3:2 ratio between authors in reviewed papers in terms of the ratio between IT specialists, authors, and medical professional authors. The ratio was higher when we observed the median number of IT specialists to medics - 3:1. There were, however, 24 papers (28%) that were authored exclusively by IT specialists. Of these, the authors of only 2 papers consulted cardiologists regarding issues related to feature extraction [112] and interpretation of the results that their model achieved [74]. This left 22 papers (25%) that were exclusively authored by non-clinical teams. These teams were prone to over-claiming and most importantly, risked producing flawed predictions. There is one study where classification criteria and the model results had not been discussed with clinicians. Despite producing a predictive model which, from a clinical perspective, incorrectly classified patients into HF severity groups, the authors still quoted that the decision tree ID3 model gave better results than all previously reported models attempting to solve the classification problem in HF (Accuracy 0.97). Incorrect classification of a patient to "at risk" group when they should have been allocated to "critically ill" because the patient was less than 42 years old and had significantly reduced left ventricular ejection fraction LVEF <40%. The patient was classified as "critically ill" when having a normal heart rate of 56 beats per minute but was aged above 60. This work demonstrates the importance of close collaboration between data analysts with domain experts in order to produce a model giving high quality, clinically appropriate and interpretable results [35]. Rammal et al. (2018) demonstrates the importance of utilising clinical expertise [32]. They used MATLAB Haar wavelets and local binary pattern (LBP) to interpret patients' chest radiographs [32]. Based on LBP assessment they classified patients into HF or non HF groups. These radiographs however were not formally reported by a radiologist consultant. Despite this serious drawback, authors claimed that their Random Forest based algorithm achieved AUC of 0.94 in the ability to classify to HF or non-HF group, which would be questionable, given that they used unvalidated set of chest radiographs to train their model [32].

6 Discussion

6.1 Engagement with domain experts

We observed that it can be easy to overstate the claims in a ML study. There was a tendency among the reviewed publications to overclaim the usefulness and applicability of the developed models to solve clinical problems. Models that are both trained and validated using retrospective observational cohort studies may be more prone to overfitting data and therefore are less able to generalise to the general population. Randomised clinical trials (RCT) remain the gold standard in research leading to introduction of new therapies and procedures in the clinical medicine. The best practice to test the applicability of the ML model in a clinical setting would be a RCT to test the ML model vs. standard care in relation to a pre-specified endpoint or outcome measure. There is a need for a close collaboration between clinicians and data scientists to prevent the production and deployment of poor models. The authors should interpret the results with a clinical team before drawing the conclusion that their model outperforms previous models in solving the classification problem. Choi et al. (2017) [39] mentioned in their conclusion, that their team "*would have benefited from well established expert medical knowledge*" such as specific features or medical ontologies when developing their predictive model for the early detection of HF. This statement left us wondering if the engagement with medical experts was an afterthought in the data mining projects, for example in instances in which the performance of the model does not meet the authors' expectations. In this case, the SVM predictive model achieved an AUC score of 0.74. We felt that Saqlain et al. (2016) exemplify the pitfalls of the clinician/data scientist knowledge gap particularly well [112]. The authors were exclusively affiliated with an IT department, yet they produced a "treatment plan for HF". On a closer look, this plan did not follow any of the current international guidelines by respected cardiology societies. Despite this, the authors claimed the model was both highly accurate and highly useful [112]. Engagement with clinical domain experts would provide greater assurance that the correct questions are asked and ensure that clinically relevant predictive models are produced. Moore et al. [96] recognised that understanding the underlying characteristics of real live clinical dataset was fundamental to enabling a critical analysis of the ML results for the sake of clinical and medical relevance.

In order to explain what measures clinicians use to evaluate whether ML application fails or not, it is important to refer to the human learning process. Clinicians, as well as other domain experts learn by experience and reflection [113]. In clinical practice, learning by experience is enabled by exposure to variety of clinical scenarios over the course of specialist training and this is followed by lifelong learning as a part of continuous professional development (CPD) and continuous medical education (CME). Clinicians continue the learning through progressive reading and reflective practice [113]. Most importantly, however, doctors use contextualisation to refine and confirm clinical diagnosis based on objective tests, prior experience and specialists' knowledge of human physiology and pathology. The skill of a quick recall of learnt facts and contextualisation allows clinicians to critically appraise the results produced by application of ML model to clinical datasets. Clinicians also reason using first principles (e.g. with understanding of fundamental biology) which is arguably very different to the machine learning paradigm.

When assessing the results of the ML experiment analysing a large dataset, clinician's first concern will be the accuracy of the classification or diagnosis and accuracy of suggested treatment. In evaluating the contributions and the impact of ML in HF management, clinicians would always like to know how relevant and applicable ML method is to the outcome of the individual patient: can the ML method produce a prediction of an event (being hospital admission, adverse reaction to treatment, worsening of HF symptoms) that can ultimately alter this individual patient's outcomes i.e. improve the quality of life, prolong the life expectancy, reduce the risk of major event like stroke or myocardial infarct? Hence, clinicians can focus on outcomes based assessment which is very important. The main benefit that clinicians would expect of having the access to accurate and reliable ML models embedded within ECR will be enhancement of safe clinical practice in line with the latest evidence-based treatments and modern diagnostic methods for the benefit of the individual patient. Involving a clinician early in the data science pipeline is critical. A clinician will typically evaluate an algorithm by benchmarking the algorithm's accuracy with the accuracy achieved by humans (e.g. consultants). This is an important benchmark that is often missed. Whilst an algorithm may show results that are statistically significant at the raw data level, if its accuracy is significantly inferior to humans' assessment, then its utility maybe called into question and lack 'clinical significance'. Moreover, clinicians will help focus data science projects on knowledge and understanding as opposed to mere accuracy measures. For example, a clinician can inspect the patient cases that were misclassified by the algorithm and using expertise to understand why those cases were misclassified.

6.2 HF prediction models in international guidelines

Despite extensive research into ML applications in HF, these ML algorithms do not yet feature strongly in international guidelines. The European Society of Cardiology (ESC), in its 85-page document with Clinical Guidelines for the Diagnosis and Treatment of Acute and Chronic Heart Failure (2016),

dedicates only a short paragraph to 'predictive models'. They state that precise risk stratification in HF remains challenging and the clinical applicability of predictive models is limited [114]. Similarly, the National Institute for Clinical Excellence (NICE) emphasize the uncertain clinical value of predictive models [6]. It should be noted, however, that several of non-ML risk models are widely used in cardiology such as GRACE [115], HEART [116], TIMI [117] and euroSCORE [118]. Several of these are focused on HF specifically. ESC HF guidelines (2016) highlighted cases in cardio-oncology when a risk score for identifying women with breast cancer may be useful. Women with breast cancer are at risk of developing HF during chemotherapy with trastuzumab and the risk score could prevent catastrophic side effects of the cardiotoxic chemotherapy [119]. In 2014 however, the ESC recommended using an online calculator to estimate a patient's 5-year risk of sudden cardiac death (SCD) due to the Hypertrophic Cardiomyopathy (HCM). The HCM Risk-SCD calculator is used frequently by cardiologists to identify patients, who are at highest risk for sudden death secondary to hypertrophic cardiomyopathy, which is one of the arrhythmogenic cardiomyopathies. The high score from HCM Risk-SCD identifies patients who would benefit most from having a prophylactic implantable cardioverter defibrillator (ICD) fitted as a primary prevention of SCD. O'Mahony, from St. Bartholomew's Centre for Inherited Cardiac Diseases, stressed that the quantification of the individual patient's risk enhances the shared decision-making process between the clinician and a patient [120]. Choosing the best treatment option 'for the patient with the patient' fulfills the ethos of 'do no harm', which is quoted as one of the most important rules for practicing clinicians by the GMC Good Clinical Practice Guide [121].

7 Gaps and Research opportunity

In this review, we identified clear gaps and areas for development in the subject of HF and data analytics. We have summarised the gaps in the literature and formulated recommendations for future work and further research within this discipline.

7.1 Clinical pathways

From a clinical perspective, one of the recurrent issues in the HF cohort is a poor uptake of evidence-based therapies. Even when effective evidence therapy is available - patients are not on optimal targeted therapies and opportunities for optimising medications while waiting for clinical reviews are missed. We are mindful of variability in access to specialists with HF expertise. To date, there has not been much consideration given to the analysis of clinical processes, clinical pathways mining and methods of monitoring patient clinical condition and medication uptitration. Despite multiple studies which consider the prediction of HF readmission to hospital, early detection of HF, there still remains an unmet need to ensure that patients are seen early by specialist teams and start lifesaving and life prolonging treatment as soon as possible.

7.2 Access to modern and divers HF databases

There have been significant improvements over the past two decades in HF pharmacotherapy and device therapy [3,114,122–125]. It is possible that models developed on historical data from 1994, for example, may have little prognostic value when applied to the patients with HF in 2021. To date, the most widely validated mortality prediction model is the Seattle Heart Failure Model (SHFM). SHFM is a mortality calculator developed on patients' data recruited to clinical trials between 1992 and 1994. SHFM is recommended by the International Society for Heart and Lung Transplantation Guidelines as a guide score used prior to left ventricle assisted device (LVAD) implantation and heart transplantation in a severe HF [126]. Simpson and McMurray [127] stated that there is a need for new models that have been designed using more contemporary cohorts of HF patients. Those models should include multiple measurements of biomarkers routinely used in clinical practice. This will allow the development of a dynamic predictive model in contrast to a model which takes into consideration only the single reading of the biomarker [127].

Another important aspect is the access to diverse HF datasets. In our review, we noted that nearly all studies utilised datasets sourced in America, Europe and Asia. There were no studies which analysed datasets from the African continent. This poses a risk of producing biased algorithms, hence we should look for diverse and highly representative modern cohorts of patients with HF.

7.3 Validation of algorithms

There seem to be frequent issues regarding validation procedures of ML models on external patient cohorts. There is a need to improve validation procedures. Validation procedures were not well described or robust enough to allow for a fair model comparison in real-world case studies. We noted that more work needs to be done with respect to the calibration, external validation, and deployment of predictive models to ensure that they are suitable for clinical use. One of the barriers or prohibiting factors may be the difficulty in getting access to external, not seen before healthcare data. Patient data are governed by data curators and access to confidential patient information is decided by regulators and ethical panels. Data governance processes are rigorous and

lengthy. Validation of the models on real data sets is however needed to support the credibility and replicability of the HF models. Once those aspects of ML application to healthcare data are addressed, there is a chance that wider clinical teams will start implementing ML models and warm up to the idea of using decision support tools (DSS) in their clinical routine. We agree with Kelly et al. (2019) [128] that in order to make fair comparisons, algorithms should be subjected to a comparison on the same external test set that is representative of the target population, using the same performance metrics. The external validation could allow clinicians to determine the best algorithm for their patients.

7.4 ML methods of the future

To be able to address the issue of model transparency and external validation, increased effort is required to develop accessible ML algorithms. The idea of using in clinical settings a method based on a "black box" mechanism will be quickly rejected by regulators and clinicians, hence there should be more focus on developing methods which explain and define how the "black box" operates. Another aspect that future methods should focus upon is the security and privacy protections of accessed data. Methods allowing data analysis at the point of data source to ensure that patients confidentiality is not breached would be an advantage. There is a need for new ML methods which allow the automation of patient data capture. ML methods that are going to be deployed and act as decision support systems should access high quality data. Automation of data capture eliminates human error and reduces the burden of administrative tasks which place a burden on the medical workforce. The time spent on data capture and ensuring high data quality is the time, that is ultimately taken away from the clinician - patient interaction. ML methods that will allow clinicians to work smarter will be in high demand, especially now, when the clinical workforce can not work any harder.

7.5 Collaboration with data curators and clinicians

Engagement with clinicians is needed at the very early stages of data analytics to minimise time spent on investigating inappropriate questions (from a clinical point of view) and to increase the utility of proposed models in the real world. Collaboration with data curators may lead to the development of data repositories that could serve as external validation sets for studies performed on different cohort of patients. Another benefit of access to population-based registries and healthcare data repositories is the opportunity to run registry based randomised controlled trials. Registry-based trials are inexpensive and less time consuming in comparison to RCT with human participants. Uncomplicated procedures around safe access to high quality data repositories will promote data driven research and prompt identification of previously undetected clinical problems.

7.6 Transparency and reporting of trials with use of ML

With an increasing number of clinical trials with ML and AI tools there is understandably an urgent need for transparent reporting of these trials. ML algorithms and validation methodologies should be carefully designed and reported to avoid research waste. CONSORT-AI and SPIRIT-AI Extension groups will address the issue of transparent and systematic reporting of trials with ML and AI [129]. In July 2019, there were 368 clinical trials in the field of AI or ML registered on ClinicalTrials.gov [129]. To date, the majority of trials were retrospective or observational, with a handful of prospective trials using AI or ML in an intervention arm [130]. We noted that the reporting of the studies varied between papers. Unfortunately, because authors of reviewed papers did not use a standard set of data and there was no unified definition for used variables, it was not possible to create a universal feature dictionary to perform further data synthesis or comprehensive meta-analysis. This highlights the importance of the use of recognised dictionaries for data collection akin to the standard sets developed by the International Consortium for Health Outcomes Measurement [131]. Health Data Research UK (HDRUK) published the 20 critical questions on transparency, replicability, ethics, and effectiveness (TREE) on best practice guidance on ML and AI research for patient benefit [132]. TREE provides a framework for researchers to inform the design, conduct, and reporting; for clinicians and policy makers, it helps to critically appraise where the new findings may be delivered for the benefit of the patient [132].

7.7 High time for RCTs

The most important gap that has not been addressed as yet is the lack of evidence that ML driven methodologies could be used in parallel with everyday standard clinical practice. Decision support tools based on predictive models, could save time and money spent in healthcare. We need robust evidence that ML methods can handle complexities of clinical reasoning before they can be safely deployed to clinical practice. Developers of predictive models should now move from the development stage to the deployment stage. As in a case of patient-specific predictions about HF readmission, they have not been widely used because of low evidence and uncertainty about the efficacy and accuracy of using measures of risk to guide clinical decisions.

Only after proving safety and positive impact on patients' outcomes could ML and AI tools be deployed to real clinical environments. The UK Medicines and Healthcare products

Regulatory Agency (MHRA) considers AI to be a medical device [133]. MHRA developed a work program to ensure that AI used for screening, diagnosis, treatment, and management of chronic conditions is treated as a medical device and is appropriately evidenced. Main areas of concern are issues of human interpretability (mentioned earlier "black box" effect and lack of transparency of AI) and adaptivity (retraining of AI models in real time). Given that AI and ML are considered a medical device, they should be tested under the same rigorous conditions as all implantable and non-implantable medical devices during prospective RCTs. Carefully designed RCTs with ML support decision tools and predictive models in an intervention arm vs standard of practice would allow objective and robust test of their effectiveness and impact on clinical practice and patients' outcomes. The next step should be careful planning of RCTs where ML guided practice could be compared to standard of care with clearly defined outcome measures like safety, improvement of diagnostic process in terms of time from presentation to diagnosis, accuracy of classification to treatment groups, improvement in time where target medication doses are achieved.

8 Conclusions

In conclusions, the choice of a 6-year window for the literature review could be considered a drawback to this study, however it should be noted that we had the intention to review ML models which were applied to contemporary cohorts of patients, who were treated with modern evidence based HF pharmacotherapy and whose clinical course was closely monitored by biomarker assays such as proBNP and Troponin [127]. Even though the review focused on studies from the last 6 years, it was observed that many research groups still used rather outdated open source cardiac dataset from UCI which was donated by Dr David W. Aha in January 1988. This database contains data from exercise stress test (EST) which has been phased out in 2016 as a first line diagnostic test for patients presenting with the new onset chest pain from the UK Clinical Guidelines number 95 published by National Institute for Clinical Excellence (NICE) [134]. Our review is limited by our literature search, where we explicitly required the search terms "heart failure", "machine learning" or "data mining" or "data analytics" or "data science" mentioned in the title or the abstract.

We have excluded studies using datasets from clinical trials. Data from clinical trials represent a highly selective cohort of the HF population. The trial data do not represent real-world healthcare data issues with missing values or an uneven distribution of features. Due to the differences in type of data presented in mobile health and telehealth datasets, we decided to exclude studies focusing on application of ML on these datasets. We recognise however the growing application of wearable technology in collecting data from ambulatory patients with HF [135]. Whilst we excluded papers that involve natural language processing (NLP) techniques as the means of interrogating HF datasets, we recognise that NLP techniques can be used to analyse clinicians' free text notes [136,137].

One of the more significant findings to emerge from this review is the fact that modern ML models have the potential to capture the complex interplay between clinical variables more effectively when compared to traditional statistical methods. This increases as the sophistication of the models increases. For example, in the extremely challenging field of natural language processing, the GPT-3 neural network has succeeded in learning latent space representations that allow it to communicate with unnerving 'human-ness'. However, the predictive power of ML models tends to correlate inversely with how explainable they are. To conclude, this presents a substantial barrier to clinical adoption and more research is needed to address the transparency and explainability of ML models. Based on our systematic literature review, we share the conclusions drawn by Di Tanna et al. (2020) [138] who concluded that despite 40 new publications on predicting risk in HF being published between 2013 and 2018, there was little evidence to show that any of 58 models described in those studies have been adopted by healthcare institutions [138]. We support their observation that there is no international or local guidance recommending one risk prediction model over another. Even when American College of Cardiology, European Society of Cardiology, or NICE guidelines mention the use of predictive models, they still claim that more research needs to be done into the clinical use of predictive models [3,6,126,139].

The authors of this review paper include clinicians with significant expertise in HF. Clinical expertise and domain knowledge facilitated the identification of inaccuracies and incorrect classifications in context of some studies presented in this paper. In addition, the findings of this review suggest that based on the ratio of clinicians to data researchers in the make-up of authors, good proportion of reviewed studies were driven by data analysts. It is important to stress that studies developing predictive models that are to be used in clinical settings should be co-driven by domain experts via a very close collaboration with data analysts. This approach will guarantee that the right research questions are asked at the right time and promote uptake.

Otherwise there is a risk of producing ML and AI algorithms which will never see the artificial light of a clinical room. In summary, we see growing potential for ML application in routine clinical practice, this however requires a shift from development stage to the deployment stage of ML models after validation in RCTs, research into clinical pathways, access to modern HF datasets and concerted effort to

improve transparency and reporting of trials with the use of ML.

Acknowledgements The authors thank Mary Rose Holman, librarian from Ulster University, for her expert assistance in conducting the systematic literature search.

Author Contributions All authors contributed to the conception and design. Literature search and data analysis were performed by A. Jasinska-Piadlo, P. Biglarbeigi and R. Bond. The first draft of the manuscript was written by A. Jasinska-Piadlo. The draft has been critically revised by R.Bond, P.Biglarbeigi, R. Brisk, P.Campbell and D. McEneaeny. All authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding Dr Jasinska-Piadlo was awarded a Doctoral Fellowship Award by Public Health Agency and Research and Development Department of the Health and Social Care in Northern Ireland, UK.

Code availability Not applicable.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Virani, S.S., Alonso, A., Benjamin, E.J., Bittencourt, M.S., Callaway, C.W., Carson, A.P., Chamberlain, A.M., Chang, A.R., Cheng, S., Delling, F.N., et al.: Heart disease and stroke statistics-2020 update: a report from the American Heart Association. Circulation **141**(9), e139 (2020)
2. Roth, G.A., Mensah, G.A., Johnson, C.O., Addolorato, G., Ammirati, E., Baddour, L.M., Barengo, N.C., Beaton, A.Z., Benjamin, E.J., Benziger, C.P., et al.: Global burden of cardiovascular diseases and risk factors, 1990–2019: update from the GBD 2019 study. J. Am. Coll. Cardiol. **76**(25), 2982 (2020)
3. O'Gara, P.T., Kushner, F.G., Ascheim, D.D., Casey, D.E., Jr., Chung, M.K., De Lemos, J.A., Ettinger, S.M., Fang, J.C., Fesmire, F.M., Franklin, B.A., et al.: 2013 ACCF/AHA guideline for the management of ST-elevation myocardial infarction: executive summary: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines. Circulation **127**(4), 529 (2013)
4. D.o.E. United Nations, S.A.P. Division. World population aging (2015). https://www.un.org/en/development/desa/population/publications/pdf/ageing/WPA2015_Report.pdf
5. B.H. Foundation. Heart failure statistics (2018). https://www.bhf.org.uk/what-we-do/our-research/heart-statistics
6. N.I. for Clinical Excellence. Chronic heart failure in adults: diagnosis and management, nice guideline [ng106] (2018). https://www.nice.org.uk/guidance/ng106
7. Jensen, P.B., Jensen, L.J., Brunak, S.: Mining electronic health records: towards better research applications and clinical care. Nat. Rev. Genet. **13**(6), 395 (2012)
8. Fry, E., and Mukherjee, S.: Tech's next big wave: Big data meets biology (2018). https://fortune.com/2018/03/19/big-data-digital-health-tech/. Access 2020 Sep 9
9. Friedman, C.P., Wong, A.K., Blumenthal, D.: Achieving a nationwide learning health system. Sci. Transl. Med. **2**(57), 57cm29 (2010)
10. Friedman, C., Rigby, M.: Conceptualising and creating a global learning health system. Int. J. Med. Informatics **82**(4), e63 (2013)
11. Szummer, K., Wallentin, L., Lindhagen, L., Alfredsson, J., Erlinge, D., Held, C., James, S., Kellerth, T., Lindahl, B., Ravn-Fischer, A., et al.: Improved outcomes in patients with ST-elevation myocardial infarction during the last 20 years are related to implementation of evidence-based treatments: experiences from the SWEDEHEART registry 1995–2014. Eur. Heart J. **38**(41), 3056 (2017)
12. Spatharou, A., Hieronimus, S., and Jenkins, J.: McKinsey: Transforming healthcare with ai: The impact on the workforce and organisations (2020). https://eithealth.eu/wp-content/uploads/2020/03/EIT-Health-and-McKinsey_Transforming-Healthcare-with-AI.pdf
13. UKGovernment: The future of healthcare: our vision for digital, data and technology in health and care.policy paper (2018). https://www.gov.uk/government/publications/the-future-of-healthcare-our-vision-for-digital-data-and-technology-in-health-and-care/the-future-of-healthcare-our-vision-for-digital-data-and-technology-in-health-and-care
14. England, H.E.: The topol programme for digital fellowships in healthcare (2019). https://topol.hee.nhs.uk/digital-fellowships/
15. Topol, E.: The topol review: Preparing the healthcare workforce to deliver the digital future (2019). https://www.hee.nhs.uk/our-work/topol-review
16. Rahimi, K., Bennett, D., Conrad, N., Williams, T.M., Basu, J., Dwight, J., Woodward, M., Patel, A., McMurray, J., MacMahon, S.: Risk prediction in patients with heart failure: a systematic review and analysis. JACC: Heart Failure **2**(5), 440 (2014)
17. Eichler, K., Zoller, M., Tschudi, P., Steurer, J.: Barriers to apply cardiovascular prediction rules in primary care: a postal survey. BMC Fam. Pract. **8**(1), 1 (2007)
18. Tripoliti, E.E., Papadopoulos, T.G., Karanasiou, G.S., Naka, K.K., Fotiadis, D.I.: Heart failure: diagnosis, severity estimation and prediction of adverse events through machine learning techniques. Comput. Struct. Biotechnol. J. **15**, 26 (2017)
19. Alba, A.C., Agoritsas, T., Jankowski, M., Courvoisier, D., Walter, S.D., Guyatt, G.H., Ross, H.J.: Risk prediction models for mortality in ambulatory patients with heart failure: a systematic review. Circulation: Heart Failure **6**(5), 881 (2013)
20. Mahajan, S.M., Heidenreich, P., Abbott, B., Newton, A., Ward, D.: Predictive models for identifying risk of readmission after index hospitalization for heart failure: a systematic review. Eur. J. Cardiovasc. Nurs. **17**(8), 675 (2018)
21. Bazoukis, G., Stavrakis, S., Zhou, J., Bollepalli, S.C., Tse, G., Zhang, Q., Singh, J.P., Armoundas, A.A.: Machine learning versus conventional clinical methods in guiding management of heart failure patients-a systematic review. Heart Fail. Rev. **26**(1), 23 (2021)

22. Qiao, N.: A systematic review on machine learning in sellar region diseases: quality and reporting items. Endocr. Connect. **8**(7), 952 (2019)
23. Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., Group, P.: Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med. **6**(7), e1000097 (2009)
24. Medline. Medline ovid database (2021). http://www.ovid.com
25. Scopus. Scopus database (2021). http://www.scopus.com
26. Somani, S., Russak, A.J., Richter, F., Zhao, S., Vaid, A., Chaudhry, F., De Freitas, J.K., Naik, N., Miotto, R., Nadkarni, G.N., et al.: Deep learning and the electrocardiogram: review of the current state-of-the-art. EP Europace (2021)
27. Benjamens, S., Dhunnoo, P., Meskó, B.: The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. NPJ Digital Med. **3**(1), 1 (2020)
28. Dey, D., Slomka, P.J., Leeson, P., Comaniciu, D., Shrestha, S., Sengupta, P.P., Marwick, T.H.: Artificial intelligence in cardiovascular imaging: JACC state-of-the-art review. J. Am. Coll. Cardiol. **73**(11), 1317 (2019)
29. Haq, A.U., Li, J.P., Memon, M.H., Nazir, S., Sun, R.: A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. Mobile Information Systems **2018** (2018)
30. Hasan, S.M.M., Mamun, M.A.A., Uddin, M.P., Hossain, M.A.: Comparative analysis of classification approaches for heart disease prediction. 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2) pp. 1–4 (2018)
31. Tiwaskar, S.A., Gosavi, R., Dubey, R., Jadhav, S., Iyer, K.: Comparison of prediction models for heart failure risk: a clinical perspective. In 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA) IEEE, pp. 1–6 (2018)
32. Rammal, H.F., Emam, A.Z.: Heart failure prediction models using big data techniques. Heart Failure **9**(5) (2018)
33. Priyanka, H., Vivek, R.: Multi model data mining approach for heart failure prediction. Int. J. Data Mining Knowl. Manage Process (IJDKP) **6**(5), 31 (2016)
34. Aljaaf, A.J., Al-Jumeily, D., Hussain, A.J., Dawson, T., Fergus, P., Al-Jumaily, M.: Predicting the likelihood of heart failure with a multi level risk assessment using decision tree. In 2015 Third International Conference on Technological Advances in Electrical, Electronics and Computer Engineering (TAEECE) (IEEE, 2015), pp. 101–106
35. Liaqat, R.M., Mehboob, B., Saqib, N.A., Khan, M.A.: Data mining approach to extract the interdependency among different attributes of cardiac patients. Int. J. Comput. Sci. Inf. Secur. **14**(7), 61 (2016)
36. Sun, J., Hu, J., Luo, D., Markatou, M., Wang, F., Edabollahi, S., Steinhubl, S.E., Daar, Z., Stewart, W.F.: Combining knowledge and data driven insights for identifying risk factors using electronic health records. In AMIA Annual Symposium Proceedings, vol. 2012 (American Medical Informatics Association), vol. 2012, p. 901 (2012)
37. Alotaibi, F.S.: Implementation of machine learning model to predict heart failure disease. Int. J. Adv. Comput. Sci. Appl. **10**(6), 261 (2019)
38. Kannan, R., Vasanthi, V.: Machine learning algorithms with ROC curve for predicting and diagnosing the heart disease. In Soft Computing and Medical Bioinformatics. Springer, pp. 63–72 (2019)
39. Choi, E., Schuetz, A., Stewart, W.F., Sun, J.: Using recurrent neural network models for early detection of heart failure onset. J. Am. Med. Inform. Assoc. **24**(2), 361 (2017)
40. Choi, E., Schuetz, A., Stewart, W.F., Sun, J.: Medical concept representation learning from electronic health records and its application on heart failure prediction. arXiv preprint arXiv:1602.03686 (2016)
41. Liang, P.Y., Wang, L.J., Wu, Y.S., Pai, T.W., Wang, C.H., Liu, M.H.: Prediction of patients with heart failure after myocardial infarction. In 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) IEEE, pp. 2009–2014 (2020)
42. Chen, Y., Qin, X., Zhang, L., Yi, B.: A Novel Method of Heart Failure Prediction Based on DPCNN-XGBOOST Model. CMC-Comput. Mater. Continua. **65**(1), 495 (2020)
43. Zhou, Y., Hou, Y., Hussain, M., Brown, S.A., Budd, T., Tang, W.W., Abraham, J., Xu, B., Shah, C., Moudgil, R., et al.: Machine Learning-Based Risk Assessment for Cancer Therapy-Related Cardiac Dysfunction in 4300 Longitudinal Oncology Patients. J. Am. Heart Assoc. **9**(23), e019628 (2020)
44. Agibetov, A., Seirer, B., Dachs, T.M., Koschutnik, M., Dalos, D., Rettl, R., Duca, F., Schrutka, L., Agis, H., Kain, R., et al.: Machine learning enables prediction of cardiac amyloidosis by routine laboratory parameters: a proof-of-concept study. J. Clin. Med. **9**(5), 1334 (2020)
45. Mathis, M.R., Engoren, M.C., Joo, H., Maile, M.D., Aaronson, K.D., Burns, M.L., Sjoding, M.W., Douville, N.J., Janda, A.M., Hu, Y., et al.: Early detection of heart failure with reduced ejection fraction using perioperative data among noncardiac surgical patients: a machine-learning approach. Anesth. Anal. **130**(5), 1188 (2020)
46. Le, M.T., Vo, M.T., Mai, L., Dao, S.V.: Predicting heart failure using deep neural network. In 2020 International Conference on Advanced Technologies for Communications (ATC) IEEE, pp. 221–225 (2020)
47. Zhang, X., Qian, B., Li, X., Wei, J., Zheng, Y., Song, L., Zheng, Q.: An interpretable fast model for predicting the risk of heart failure. In Proceedings of the 2019 SIAM International Conference on Data Mining (SIAM), pp. 576–584 (2019)
48. Austin, P.C., Tu, J.V., Ho, J.E., Levy, D., Lee, D.S.: Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. J. Clin. Epidemiol. **66**(4), 398 (2013)
49. Garg, R., Dong, S., Shah, S., Jonnalagadda, S.R.: A bootstrap machine learning approach to identify rare disease patients from electronic health records. arXiv preprint arXiv:1609.01586 (2016)
50. Escamilla, A.K.G., El Hassani, A.H., Andres, E.: Dimensionality Reduction in Supervised Models-based for Heart Failure Prediction (2019)
51. Africa, A.: A rough set-based data model for heart disease diagnostics. ARPN J. Eng. Appl. Sci. **11**(15), 9350 (2016)
52. Le, H.M., Tran, T.D., Van Tran, L.: Automatic heart disease prediction using feature selection and data mining technique. J. Comput. Sci. Cybern. **34**(1), 33 (2018)
53. Rehman, A., Khan, A., Ali, M.A., Khan, M.U., Khan, S.U., Ali, L.: Performance Analysis of PCA, Sparse PCA, Kernel PCA and Incremental PCA Algorithms for Heart Failure Prediction. in *2020 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)* (2020), pp. 1–5. https://doi.org/10.1109/ICECCE49384.2020.9179199
54. Taslimitehrani, V., Dong, G., Pereira, N.L., Panahiazar, M., Pathak, J.: Developing EHR-driven heart failure risk prediction models using CPXR (Log) with the probabilistic loss function. J. Biomed. Inform. **60**, 260 (2016)
55. Ahmad, T., Lund, L.H., Rao, P., Ghosh, R., Warier, P., Vaccaro, B., Dahlström, U., O'connor, C.M., Felker, G.M., Desai, N.R.: Machine learning methods improve prognostication, identify clinically distinct phenotypes, and detect heterogeneity in response to therapy in a large cohort of heart failure patients. J. Am. Heart Assoc. **7**(8), e008081 (2018)
56. Ben-Assuli, O., Heart, T., Shlomo, N., Klempfner, R.: Bringing big data analytics closer to practice: A methodological explana-

tion and demonstration of classification algorithms. Health Policy Technol. **8**(1), 7 (2019)
57. Kwon, J.m., Kim, K.H., Jeon, K.H., Lee, S.E., Lee, H.Y., Cho, H.J., Choi, J.O., Jeon, E.S., Kim, M.S., Kim, J.J., et al.: Artificial intelligence algorithm for predicting mortality of patients with acute heart failure. PloS one **14**(7), e0219302 (2019)
58. Suzuki, S., Yamashita, T., Sakama, T., Arita, T., Yagi, N., Otsuka, T., Semba, H., Kano, H., Matsuno, S., Kato, Y., et al.: Comparison of risk models for mortality and cardiovascular events between machine learning and conventional logistic regression analysis. PLoS ONE **14**(9), e0221911 (2019)
59. Adler, E.D., Voors, A.A., Klein, L., Macheret, F., Braun, O.O., Urey, M.A., Zhu, W., Sama, I., Tadel, M., Campagnari, C., et al.: Improving risk prediction in heart failure using machine learning. Eur. J. Heart Fail. **22**(1), 139 (2020)
60. Wang, Z., Chen, L., Zhang, J., Yin, Y., Li, D.: Multi-view ensemble learning with empirical kernel for heart failure mortality prediction. Int. J. Numer. Methods Biomed. Eng. **36**(1), e3273 (2020)
61. Gong, J., Bai, X., Li, D.a., Zhao, J., Li, X.: Prognosis analysis of heart failure based on recurrent attention model. IRBM **41**(2), 71 (2020)
62. Panahiazar, M., Taslimitehrani, V., Pereira, N., Pathak, J.: Using EHRs and machine learning for heart failure survival analysis. Stud. Health Technol. Inform. **216**, 40 (2015)
63. Kubus, L., Yastrebov, A., Poczeta, K., Poterala, M., Gromadzinski, L.: The use of fuzzy cognitive maps in evaluation of prognosis of chronic heart failure patients. In 2018 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA) IEEE, pp. 191–196 (2018)
64. Chicco, D., Jurman, G.: Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Med. Inform. Decis. Mak. **20**(1), 16 (2020)
65. Sax, D.R., Mark, D.G., Huang, J., Sofrygin, O., Rana, J.S., Collins, S.P., Storrow, A.B., Liu, D., Reed, M.E.: Use of machine learning to develop a risk-stratification tool for emergency department patients with acute heart failure. Ann. Emerg. Med. **77**(2), 237 (2021)
66. Dziewięcka, E., Gliniak, M., Winiarczyk, M., Karapetyan, A., Wiśniowska-Śmiałek, S., Karabinowska, A., Dziewięcki, M., Podolec, P., Rubiś, P.: Mortality risk in dilated cardiomyopathy: the accuracy of heart failure prognostic models and dilated cardiomyopathy-tailored prognostic model. ESC Heart Failure **7**(5), 2455 (2020)
67. Wang, Z., Yao, L., Li, D., Ruan, T., Liu, M., Gao, J.: Mortality prediction system for heart failure with orthogonal relief and dynamic radius means. Int. J. Med. Informatics **115**, 10 (2018)
68. Chu, J., Dong, W., Huang, Z.: Endpoint prediction of heart failure using electronic health records. J. Biomed. Inform. **109**, 103518 (2020). https://doi.org/10.1016/j.jbi.2020.103518
69. Jing, L., Cerna, A.E. Ulloa, Good, C.W., Sauers, N.M., Schneider, G., Hartzel, D.N., Leader, J.B., Kirchner, H.L., Hu, Y., Riviello, D.M. et al.: A machine learning approach to management of heart failure populations. Heart Failure **8**(7), 578 (2020)
70. Javeed, A., Rizvi, S.S., Zhou, S., Riaz, R., Khan, S.U., Kwon, S.J.: Heart risk failure prediction using a novel feature selection method for feature refinement and neural network for classification. Mobile Information Systems **2020** (2020)
71. Stampehl, M., Friedman, H.S., Navaratnam, P., Russo, P., Park, S., Obi, E.N.: Risk assessment of post-discharge mortality among recently hospitalized Medicare heart failure patients with reduced or preserved ejection fraction. Curr. Med. Res. Opin. **36**(2), 179 (2020)
72. Tse, G., Zhou, J., Woo, S.W.D., Ko, C.H., Lai, R.W.C., Liu, T., Liu, Y., Leung, K.S.K., Li, A., Lee, S., et al.: Multi-modality machine learning approach for risk stratification in heart failure with left ventricular ejection fraction ≤45%. ESC Heart Failure **7**(6), 3716 (2020)
73. Blackstone, E.H., Rajeswaran, J., Cruz, V.B., Hsich, E.M., Koprivanac, M., Smedira, N.G., Hoercher, K.J., Thuita, L., Starling, R.C.: Continuously updated estimation of heart transplant waitlist mortality. J. Am. Coll. Cardiol. **72**(6), 650 (2018)
74. Liu, R., Zolfaghar, K., Chin, S.c., Roy, S.B., Teredesai, A.: A framework to recommend interventions for 30-day heart failure readmission risk. In 2014 IEEE International Conference on Data Mining IEEE, pp. 911–916 (2014)
75. Lorenzoni, G., Sabato, S.S., Lanera, C., Bottigliengo, D., Minto, C., Ocagli, H., De Paolis, P., Gregori, D., Iliceto, S., Pisanò, F.: Comparison of machine learning techniques for prediction of hospitalization in heart failure patients. J. Clin. Med. **8**(9), 1298 (2019)
76. Cheung, B.L.P., Dahl, D.: Deep learning from electronic medical records using attention-based cross-modal convolutional neural networks. In 2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI) IEEE, pp. 222–225 (2018)
77. Mahajan, S.M., Mahajan, A.S., King, R., Negahban, S.: Predicting risk of 30-day readmissions using two emerging machine learning methods. In Nursing Informatics 2018. IOS Press, pp. 250–255 (2018)
78. Mahajan, S.M., Burman, P., Newton, A., Heidenreich, P.A.: A validated risk model for 30-day readmission for heart failure. In MEDINFO 2017: Precision Healthcare Through Informatics. IOS Press, pp. 506–510 (2017)
79. Frizzell, J.D., Liang, L., Schulte, P.J., Yancy, C.W., Heidenreich, P.A., Hernandez, A.F., Bhatt, D.L., Fonarow, G.C., Laskey, W.K.: Prediction of 30-day all-cause readmissions in patients hospitalized for heart failure: comparison of machine learning and other statistical approaches. JAMA Cardiol. **2**(2), 204 (2017)
80. Mahajan, S.M., Ghani, R.: Using ensemble machine learning methods for predicting risk of readmission for heart failure. In MedInfo, pp. 243–247 (2019)
81. Shameer, K., Johnson, K.W., Yahi, A., Miotto, R., Li, L., Ricks, D., Jebakaran, J., Kovatch, P., Sengupta, P.P., Gelijns, S., et al.: Predictive modeling of hospital readmission rates using electronic medical record-wide machine learning: a case-study using Mount Sinai heart failure cohort. In Pacific Symposium on Biocomputing 2017. World Scientific, pp. 276–287 (2017)
82. Xiao, C., Ma, T., Dieng, A.B., Blei, D.M., Wang, F.: Readmission prediction via deep contextual embedding of clinical concepts. PLoS ONE **13**(4), e0195024 (2018)
83. Jiang, W., Siddiqui, S., Barnes, S., Barouch, L.A., Korley, F., Martinez, D.A., Toerper, M., Cabral, S., Hamrock, E., Levin, S.: Readmission risk trajectories for patients with heart failure using a dynamic prediction approach: retrospective study. JMIR Medical Informatics **7**(4) (2019)
84. Sideris, C., Pourhomayoun, M., Kalantarian, H., Sarrafzadeh, M.: A flexible data-driven comorbidity feature extraction framework. Comput. Biol. Med. **73**, 165 (2016)
85. Sarijaloo, F., Park, J., Zhong, X., Wokhlu, A.: Predicting 90 day acute heart failure readmission and death using machine learning-supported decision analysis. Clinical Cardiology (2020)
86. Ben-Assuli, O., Heart, T., Vest, J.R., Ramon-Gonen, R., Shlomo, N., Klempfner, R.: Profiling readmissions using hidden markov model-the case of congestive heart failure. Information Systems Management pp. 1–13 (2020)
87. Lewis, G.E. Maor, Beladev, M., Maor, G., Radinsky, K., Hermann, D., Litani, Y., Geller, T., Pines, J.M., et al.: Comparison of deep learning with traditional models to predict preventable acute care use and spending among heart failure patients. Scientific Reports **11**

88. Lu, X.H., Liu, A., Fuh, S.C., Lian, Y., Guo, L., Yang, Y., Marelli, A., Li, Y.: Recurrent disease progression networks for modelling risk trajectory of heart failure. PLoS ONE **16**(1), e0245177 (2021)
89. Savitz, S., Leong, T., Sung, S., Lee, K., Rana, J., Tabada, G., Go, A.: Novel Data Domains and Machine Learning Modestly Improved Performance of Risk Calculators for Heart Failure Readmission. Health Serv. Res. **55**, 85 (2020)
90. Hu, Z., Du, D.: A new analytical framework for missing data imputation and classification with uncertainty: Missing data imputation and heart failure readmission prediction. PLoS ONE **15**(9), e0237724 (2020)
91. Liu, D., Lei, L., Ruan, T., He, P.: Constructing large scale cohort for clinical study on heart failure with electronic health record in regional healthcare platform: challenges and strategies in data reuse. Chin. Med. Sci. J. **34**(2), 90 (2019)
92. Rjeily, C.B., Badr, G., Al Hassani, A.H., Andres, E.: Predicting heart failure class using a sequence prediction algorithm. In 2017 Fourth International Conference on Advances in Biomedical Engineering (ICABME) IEEE, pp. 1–4 (2017)
93. Yuan, Y.B., Qiu, W.Q., Wang, Y.J., Gao, J., He, P.: Classification of heart failure with polynomial smooth support vector machine. In 2017 International Conference on Machine Learning and Cybernetics (ICMLC), vol. 1 . IEEE, pp. 48–54 (2017)
94. Balabaeva, K., Kovalchuk, S.V., Metsker, O.G.: Dynamic features impact on the quality of chronic heart failure predictive modelling. In pHealth, pp. 179–184 (2019)
95. Saqlain, M., Hussain, W., Saqib, N.A., Khan, M.A.: Identification of heart failure by using unstructured data of cardiac patients. In 2016 45th International Conference on Parallel Processing Workshops (ICPPW). IEEE, pp. 426–431 (2016)
96. Moore, L., Kambhampati, C., Cleland, J.G.: Classification of a real live heart failure clinical dataset-Is TAN Bayes better than other Bayes?. In 2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE, pp. 882–887 (2014)
97. Hussain, L., Lone, K.J., Awan, I.A., Abbasi, A.A., Pirzada, J.u.R.: Detecting congestive heart failure by extracting multimodal features with synthetic minority oversampling technique (SMOTE) for imbalanced data using robust machine learning techniques. Waves in Random and Complex Media pp. 1–24 (2020)
98. Nagamine, T., Gillette, B., Pakhomov, A., Kahoun, J., Mayer, H., Burghaus, R., Lippert, J., Saxena, M.: Multiscale classification of heart failure phenotypes by unsupervised clustering of unstructured electronic medical record data. Sci. Rep. **10**(1), 1 (2020)
99. Gu, J., Pan, J.A., Lin, H., Zhang, J.F., Wang, C.Q.: Characteristics, prognosis and treatment response in distinct phenogroups of heart failure with preserved ejection fraction. International Journal of Cardiology **323**, 148 (2021)
100. Schrub, F., Oger, E., Bidaut, A., Hage, C., Charton, M., Daubert, J.C., Leclercq, C., Linde, C., Lund, L., Donal, E.: Heart failure with preserved ejection fraction: A clustering approach to a heterogenous syndrome. Arch. Cardiovasc. Dis. **113**(6–7), 381 (2020)
101. Kaptein, Y.E., Karagodin, I., Zuo, H., Lu, Y., Zhang, J., Kaptein, J.S., Strande, J.L.: Identifying Phenogroups in patients with subclinical diastolic dysfunction using unsupervised statistical learning. BMC Cardiovasc. Disord. **20**(1), 1 (2020)
102. Hedman, Å.K., Hage, C., Sharma, A., Brosnan, M.J., Buckbinder, L., Gan, L.M., Shah, S.J., Linde, C.M., Donal, E., Daubert, J.C., et al.: Identification of novel pheno-groups in heart failure with preserved ejection fraction using machine learning. Heart **106**(5), 342 (2020)
103. Chen, P., Dong, W., Lu, X., Kaymak, U., He, K., Huang, Z.: Deep representation learning for individualized treatment effect estimation using electronic health records. J. Biomed. Informatics (2019). https://doi.org/10.1016/j.jbi.2019.103303
104. Balabaeva, K., Kovalchuk, S.: Comparison of temporal and non-temporal features effect on machine learning models quality and interpretability for chronic heart failure patients. Procedia Comput. Sci. **156**, 87 (2019)
105. Kumar, G.K.: An optimized particle swarm optimization based ANN Model for clinical disease prediction. Indian J. Sci. Technol. **9** (2016)
106. Nouraei, H., Rabkin, S.W.: A new approach to the clinical subclassification of heart failure with preserved ejection fraction. Int. J. Cardiol. **331**, 138 (2021)
107. Ali, L., Niamat, A., Khan, J.A., Golilarz, N.A., Xingzhong, X., Noor, A., Nour, R., Bukhari, S.A.C.: An optimized stacked support vector machines based expert system for the effective prediction of heart failure. IEEE Access **7**, 54007 (2019)
108. Dua, D., Graff, C.: UCI machine learning repository (2017). http://archive.ics.uci.edu/ml
109. Aleryani, A., Wang, W., De La Iglesia, B.: Multiple imputation ensembles (MIE) for dealing with missing data. SN Comput. Sci. **1**(3), 1 (2020)
110. Hicks, S.A., Isaksen, J.L., Thambawita, V., Ghouse, J., Ahlberg, G., Linneberg, A., Grarup, N., Strümke, I., Ellervik, C., Olesen, M.S., et al.: Explaining deep neural networks for knowledge discovery in electrocardiogram analysis. Sci. Rep. **11**(1), 1 (2021)
111. Janssen, K.J., Donders, A.R.T., Harrell, F.E., Jr., Vergouwe, Y., Chen, Q., Grobbee, D.E., Moons, K.G.: Missing covariate data in medical research: to impute is better than to ignore. J. Clin. Epidemiol. **63**(7), 721 (2010)
112. Saqlain, M., Athar, A., Saqib, N.A., Khan, M.A.: Developing a Classification Model for an Effective Treatment of Heart Failure. Int. J. Comput. Sci. Inform. Secur. **14**(8), 413 (2016)
113. Slotnick, H.: How doctors learn: the role of clinical problems across the medical school-to-practice continuum. Acad. Med.: J. Assoc. Am. Med. Colleges **71**(1), 28 (1996)
114. Ponikowski, P., Voors, A.A., Anker, S.D., Bueno, H., Cleland, J.G., Coats, A.J., Falk, V., González-Juanatey, J.R., Harjola, V.P., Jankowska, E.A., et al.: ESC Scientific Document Group. 2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: The Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC) Developed with the special contribution of the Heart Failure Association (HFA) of the ESC. Eur Heart J **37**(27), 2129 (2016)
115. Bradshaw, P.J., Ko, D.T., Newman, A.M., Donovan, L.R., Tu, J.V.: Validity of the GRACE (Global Registry of Acute Coronary Events) acute coronary syndrome prediction model for six month post-discharge death in an independent data set. Heart **92**(7), 905 (2006) https://doi.org/10.1136/hrt.2005.073122. https://heart.bmj.com/content/92/7/905
116. Frisoli, T.M., Nowak, R., Evans, K.L., Harrison, M., Alani, M., Varghese, S., Rahman, M., Noll, S., Flannery, K.R., Michaels, A., et al.: Henry Ford HEART score randomized trial: rapid discharge of patients evaluated for possible myocardial infarction. Circulation: Cardiovascular Quality and Outcomes **10**(10), e003617 (2017)
117. Antman, E.M., Cohen, M., Bernink, P.J.L.M., McCabe, C.H., Horacek, T., Papuchis, G., Mautner, B., Corbalan, R., Radley, D., Braunwald, E.: The TIMI risk score for unstable angina/non-ST elevation MIA method for prognostication and therapeutic decision making. JAMA **284**(7), 835 (2000). https://doi.org/10.1001/jama.284.7.835
118. Nashef, S.A., Roques, F., Hammill, B.G., Peterson, E.D., Michel, P., Grover, F.L., Wyse, R.K., Ferguson, T.B.: Validation of European system for cardiac operative risk evaluation (EuroSCORE) in North American cardiac surgery. Eur. J. Cardiothorac. Surg. **22**(1), 101 (2002)

119. Ezaz, G., Long, J.B., Gross, C.P., Chen, J.: Risk prediction model for heart failure and cardiomyopathy after adjuvant trastuzumab therapy for breast cancer. J. Am. Heart Assoc. **3**(1), e000472 (2014)
120. ESC: Preventing sudden death in hypertrophic cardiomyopathy: new backing for esc guidelines (hcm-evidence) (2017). https://www.escardio.org/The-ESC/Press-Office/Press-releases/preventing/-sudden/-death/-in/-hypertrophic/-cardiomyopathy/-new/-backing/-for/-esc/-guidelines/-hcm/-evidence
121. Council, G.M.: Good medical practice (2013). ISBN: 978-0-901458-73-5
122. McMurray, J.J., Packer, M., Desai, A.S., Gong, J., Lefkowitz, M.P., Rizkala, A.R., Rouleau, J.L., Shi, V.C., Solomon, S.D., Swedberg, K., et al.: Angiotensin-neprilysin inhibition versus enalapril in heart failure. N. Engl. J. Med. **371**, 993 (2014)
123. McMurray, J.J., Solomon, S.D., Inzucchi, S.E., Køber, L., Kosiborod, M.N., Martinez, F.A., Ponikowski, P., Sabatine, M.S., Anand, I.S., Bělohlávek, J., et al.: Dapagliflozin in patients with heart failure and reduced ejection fraction. N. Engl. J. Med. **381**(21), 1995 (2019)
124. Cleland, J.G., Daubert, J.C., Erdmann, E., Freemantle, N., Gras, D., Kappenberger, L., Tavazzi, L.: Longer-term effects of cardiac resynchronization therapy on mortality in heart failure [the CArdiac REsynchronization-Heart Failure (CARE-HF) trial extension phase]. Eur. Heart J. **27**(16), 1928 (2006)
125. Elming, M.B., Nielsen, J.C., Haarbo, J., Videbæk, L., Korup, E., Signorovitch, J., Olesen, L.L., Hildebrandt, P., Steffensen, F.H., Bruun, N.E., et al.: Age and outcomes of primary prevention implantable cardioverter-defibrillators in patients with nonischemic systolic heart failure. Circulation **136**(19), 1772 (2017)
126. Feldman, D., Pamboukian, S.V., Teuteberg, J.J., Birks, E., Lietz, K., Moore, S.A., Morgan, J.A., Arabia, F., Bauman, M.E., Buchholz, H.W., et al.: The 2013 International Society for Heart and Lung Transplantation Guidelines for mechanical circulatory support: executive summary. J. Heart Lung Transplant. **32**(2), 157 (2013)
127. Simpson, J., McMurray, J.J.: Prognostic modeling in heart failure: time for a reboot. Prognostic modeling in heart failure: time for a reboot (2018)
128. Kelly, C.J., Karthikesalingam, A., Suleyman, M., Corrado, G., King, D.: Key challenges for delivering clinical impact with artificial intelligence. BMC Med. **17**(1), 1 (2019)
129. Liu, X., Faes, L., Calvert, M.J., Denniston, A.K.: Extension of the CONSORT and SPIRIT statements. Lancet **394**(10205), 1225 (2019)
130. Topol, E.: *Deep medicine: how artificial intelligence can make healthcare human again* (Hachette UK, 2019)
131. Burns, D.J., Arora, J., Okunade, O., Beltrame, J.F., Bernardez-Pereira, S., Crespo-Leiro, M.G., Filippatos, G.S., Hardman, S., Hoes, A.W., Hutchison, S., et al.: International consortium for health outcomes measurement (ICHOM): standardized patient-centered outcomes measurement set for heart failure patients. Heart Failure **8**(3), 212 (2020)
132. Vollmer, S., Mateen, B.A., Bohner, G., Király, F.J., Ghani, R., Jonsson, P., Cumbers, S., Jonas, A., McAllister, K.S., Myles, P., et al.: Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. bmj **368** (2020)
133. UKGovernment. Software and ai as a medical device change programme (2021). https://www.gov.uk/government/publications/software-and-ai-as-a-medical-device-change-programme/software-and-ai-as-a-medical-device-change-programme
134. N.I. for Clinical Excellence. Recent-onset chest pain of suspected cardiac origin: assessment and diagnosis clinical guideline [cg95] (2016)
135. Barrett, M., Boyne, J., Brandts, J., Brunner-La Rocca, H.P., De Maesschalck, L., De Wit, K., Dixon, L., Eurlings, C., Fitzsimons, D., Golubnitschaja, O., et al.: Artificial intelligence supported patient self-care in chronic heart failure: a paradigm shift from reactive to predictive, preventive and personalised care. Epma J. **10**(4), 445 (2019)
136. Topaz, M., Radhakrishnan, K.: Suzanne Blackley2, Victor Lei2, Kenneth Lai4, and Li Zhou 1, 2, 4. West. J. Nurs. Res. **1**, 19 (2016)
137. Heidenreich, P.A.: Can natural language processing fulfill the promise of electronic medical records? J. Cardiac Fail. **20**(7), 465 (2014)
138. Di Tanna, G.L., Wirtz, H., Burrows, K.L., Globe, G.: Evaluating risk prediction models for adults with heart failure: A systematic literature review. PLoS ONE **15**(1), e0224135 (2020)
139. Members, A.F., Dickstein, K., Cohen-Solal, A., Filippatos, G., McMurray, J.J., Ponikowski, P., Poole-Wilson, P.A., Strömberg, A., van Veldhuisen, D.J., Atar, D., et al.: ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure 2008: the Task Force for the Diagnosis and Treatment of Acute and Chronic Heart Failure 2008 of the European Society of Cardiology. Developed in collaboration with the Heart Failure Association of the ESC (HFA) and endorsed by the European Society of Intensive Care Medicine (ESICM). European Heart Journal **29**(19), 2388 (2008)